Lab08

Motif-Based Analysis.

Transcription factors are proteins that control gene expression - the degree to which specific genes are turned “on” or “off” - by binding to nearby DNA. Each transcription factor recognizes and binds to a specific sequence in the DNA alphabet (A, C, G, and T) known as a consensus site.
MEME-ChIP1 is a web-based tool for motif-based sequence analysis of large-scale DNA or RNA data sets. It provides computationally efficient algorithms for discovering and analyzing the sequence motifs characteristic of transcription factor (TF) binding sites, RNA protein binding sites and promoter elements. MEME-ChIP performs:

- de novo motif discovery
- motif enrichment analysis
- motif location analysis
- motif clustering

A few web links:

- https://meme-suite.org/meme/index.html
- https://meme-suite.org/meme/tools/meme-chip
- https://jaspar.elixir.no/
- https://www.encodeproject.org/

Document all work during the dry lab for each exercise, defining all your tools and input parameters, data output, and interpretation.

From Lab08, there are two Assignments to be submitted to myCourses:
Discussion 8.1
Activity 8.1

Motif-based analysis of ChIP-seq data

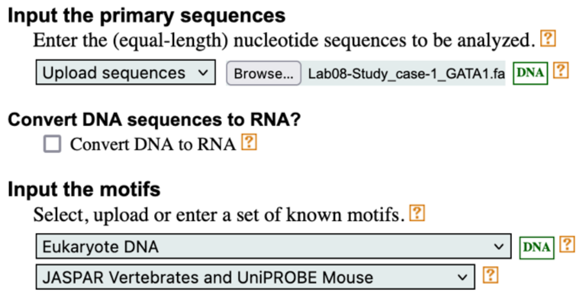
Well-established protocols exist for identifying the peaks of protein binding in ChIP-seq experiments. After the initial experimental output, we need to convert a file containing the genomic locations of ChIP-seq peak summits into DNA sequences in FASTA format (this is already done for this activity).

After that, running MEME-ChIP requires accessing the MEME-ChIP input form over the web, uploading the FASTA file of sequences, clicking a button if single-stranded analysis (for

RNA) is desired, choosing a database of known motifs from a menu and clicking the 'submit' button.

Study Case 1: GATA1 ChIP-seq in human PDBE cells. In this study case, it is demonstrated the motif-based analysis of ChIP-seq data by using GATA1 ChIP-seq in human peripheral blood-derived erythroblast (PBDE) cells produced by the Encyclopedia of DNA Elements (ENCODE) Consortium. GATA1 is a transcriptional activator, which serves as a general switching factor for erythroid development. GATA1 is known to bind to DNA sites with the consensus sequence **'[AT]GATA[AG]'** within regulatory regions of globin genes and other genes expressed in erythroid cells.

Use the file Lab08-Study_case-1_GATA1.fa in myCourses > Content > Labs with 2,727 GATA1 ChIP-seq peaks predicted by the Farnham laboratory at the University of Southern California (USC). These peaks range from 213 bp to 9,888 bp in length, with a mean length of 967.6 bp. Each peak has been replaced with the 500-bp genomic region around its center. In MEME-ChIP website, use the fasta file (with 500-bp regions) as input for motif enrichment analysis ("Input the primary sequences"). Under 'Select, upload or enter a set of known motifs' use **Eukaryote DNA** and the 'JASPAR Vertebrates and UniPROBE Mouse' **database** option.



Input the primary sequences

Enter the (equal-length) nucleotide sequences to be analyzed. ?

Upload sequences ⌄ Browse... Lab08-Study_case-1_GATA1.fa DNA ?

Convert DNA sequences to RNA?

□ Convert DNA to RNA ?

Input the motifs

Select, upload or enter a set of known motifs. ?

Eukaryote DNA ⌄ DNA ?

JASPAR Vertebrates and UniPROBE Mouse ⌄ ?

Hint: To view detailed descriptions of all the available motif databases, click the help menu ('?').

MEME-ChIP requires the input sequences to be provided by using the DNA alphabet (e.g., ACGTN). If you are analyzing RNA sequences (e.g., from a CLIP-seq experiment), you need to first convert them to DNA sequences (replacing U by T).

Provide your email address. Under 'Input job details', enter your email address. This data analysis in the MEME-ChIP server will take some time to process (1-2 hours). We will launch the run, but we'll see the analysis report that has already been processed.

Although most parameters are optional, we will review them during the discussion.

- Universal options
- MEME options
- ~~STREME options~~
- CentriMo options

Submit your analysis once all is ready. At the new tab in your browser (*Please wait. Your MEME-ChIP job is now queued awaiting available resources*), save the link address (https://meme-suite.org...), which corresponds to your specific job. Although the server asks you for your e-mail, I have found that it is common not to receive any results in your inbox (RIT spam filters?).

Once you receive your results, you can save them locally. For today, view the MEME-CHIP from the saved report (Lab08-Study_case_1) available in MyCourses. Open the file ***index.html*** on your laptop.

- The motifs are listed in order of statistical significance.
- For each motif, MEME-ChIP displays its ***logo***, the program that reported it, its statistical significance (as estimated by that program), links to similar known motifs and the positional distribution of the motif in the sequences.
- For convenience, MEME-ChIP groups motifs that are very similar to each other and display only the most significant ones. Clicking on the ‘Show 6 more’ link under the top-scoring motif displays all seven in the group.
- Clicking on ‘CentriMo’ will automatically select similar motifs and display them via the CentriMo output.
- Clicking on ‘MEME’ in the ‘Discovery/Enrichment Program’ column will take you to the output of MEME (motif discovery).
- We won’t cover DREME / STREME outputs this time.
- If the motif was detected by CentriMo, the ‘Known or Similar Motifs’ column contains a link to the motif database entry for that known motif. This is the case with the most significant motif found in the GATA1 ChIP-seq peaks (‘Tal1::Gata1’), which is believed to be bound by GATA1 in complex with Tal1 (another DNA-binding complex). The extremely centered and symmetrical motif distribution plot for this known motif for the ChIP-ed TF indicates that the ChIP-seq protocol and peak-calling pipeline were highly successful in this experiment.
- Additionally, search for similar motifs in the JASPAR database. You will need to search by the MA profile ID. Go to https://jaspar.elixir.no and search for MA0140.1.
- If the motif was detected by MEME (denovo motif discovery), it will give you the regular motif expression, sequence logo, and sequence position sites from the input.

Discussion 8.1

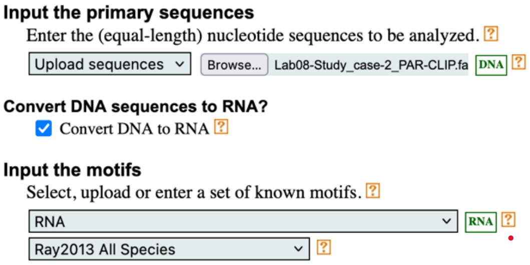
Write down the steps and parameters (methods) you used, describing the components and what they represent. Focus on the most significant motif and the 2-3 most similar significant motifs and explore known or similar motifs in the JASPAR database. Once you finish, submit the report with explanations to myCourses (in Assignments).

Motif analysis of PAR-CLIP cross-linking sites

We will look into a family of RNA-binding proteins, which are evolutionarily highly conserved from yeast to humans. The proteins in this family have sequence-specific RNA-binding domains located in their C termini that frequently bind the 3' UTR of target mRNAs.

We will analyze the 1,236 cross-linking sites identified by using conventional PAR-CLIP (A Method for Transcriptome-Wide Identification of RNA Binding Protein Interaction Sites). Instead of 500-bp, a selection of 100-bp for each cross-linking site with the genomic region surrounding its center is provided for motif enrichment.

In MEME-ChIP, we'll use the RNA (Ray 2013 database – All) motif database, which contains RNA motifs from diverse eukaryotes. Make sure you select "Convert DNA to RNA". Check your parameters and click submit.



Input the primary sequences

Enter the (equal-length) nucleotide sequences to be analyzed. ?

Upload sequences ˅ Browse... Lab08-Study_case-2_PAR-CLIP.fa DNA ?

Convert DNA sequences to RNA?

☑ Convert DNA to RNA ?

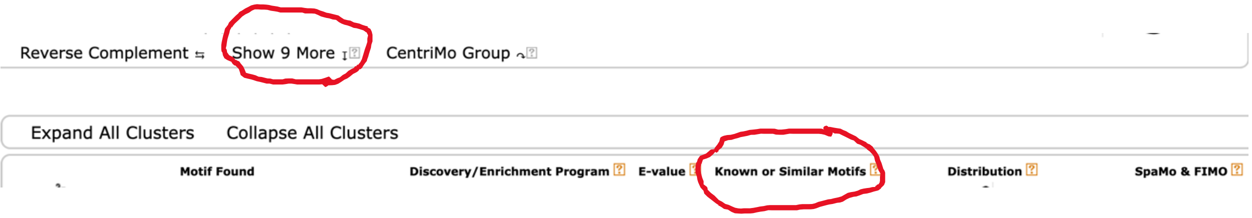
Input the motifs

Select, upload or enter a set of known motifs. ?

RNA ˅ RNA ?

Ray2013 All Species ˅ ?

Double-check your options and submit your job. In your results, what is the **most significant motif** found in the 1,236 RNA-protein crosslinking regions that generally matches known RNA-binding motifs? How many "**more" similar motifs** are detected? What protein family is associated with "known or similar motifs"?



Hint: search for "IPR001313" in InterPro, a protein family classification database: https://www.ebi.ac.uk/interpro/

Activity 8.1

Once you finish, submit the report to myCourses (in Assignments). You only need to screenshot the top four motifs from the main MEME-ChIP output and answer the above questions. Please write down the steps and parameters (methods) you used.