

PRACTICAL: MOTIF ANALYSIS ON CHIP-SEQ PEAKS

Practical: motif analysis on ChIP-seq peaks	1
INTRODUCTION	2
<i>Peak sizes</i>	2
TUTORIAL	3
<i>Organization of the Regulatory Sequence Analysis Web site</i>	3
<i>Fetching sequences from UCSC to RSAT</i>	3
<i>Discovering motifs in a single ChIP-seq peak set</i>	4
<i>Questions</i>	4
EXERCISES	5
<i>Differential analysis</i>	5
<i>Negative controls</i>	5
<i>MEME chips</i>	5
FURTHER READING	6

INTRODUCTION

The goal of this tutorial is to apply various algorithms in order to discover transcription factor binding motifs from ChIP-seq peaks.

The study case is based on the following publication:

- Theodorou, V., Stark, R., Menon, S. & Carroll, J. S. (2013). GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. *Genome Res* 23,12-22.
Pubmed [[23172872](#)]; GEO series [[GSE40129](#)]

The authors ran ChIP-seq experiments to characterize the binding locations of the human transcription factor ESR1, and studied the interactions with two related factors: GATA3 and Fox1.

We combined the MACS and PeakSplitter tools to identify genomic locations bound by the E2-activated Oestrogen Receptor (ER) in the breast cancer cell line MCF-7, under two conditions:

- inhibition of the Gata3 factor by small interference RNA (files with prefix siGATA in the Table below)
- no treatment (files with prefix siNT)

Peak regions were obtained by running the peak calling program MACS on the aligned reads, and the multi-peak regions were further split into individual peaks.

To access the peak sets, click the following links in the left frame of the course supporting web site.

ChIP-seq -> Datasets -> Peaks for motif analysis

Peak sizes

File	Peak number	Sum of peak sizes (bp)
siGATA_ER_E2_r3_MACS_PeakSplitter.bed	7.041	1.476.646
siGATA_ER_E2_r3_MACS_peaks.bed	5.134	1.966.163
siNT_ER_E2_r3_MACS_PeakSplitter.bed	6.991	1.801.622
siNT_ER_E2_r3_MACS_peaks.bed	5.533	2.201.492

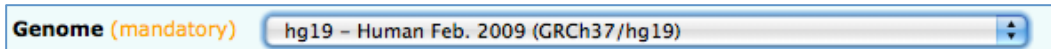
TUTORIAL

Organization of the Regulatory Sequence Analysis Web site

1. Open a connection to a Regulatory Sequence Analysis Tools server. You can choose between various instances.
 - Server at Roscoff (recommended for this course) <http://rsat.sb-roscoff.fr/>
 - Main server (currently in Brussels) <http://www.rsat.eu/>
2. In the tool list, click **Doc and help -> Tutorials**.
 - The flow chart provides a simplified view of the tools and their possible inter-connections.
 - Below this flow chart, you will find a list of links to tutorial page, explaining the main concepts and putting them in practice on selected case studies.
3. Detailed information about the tool can be found in the publication list, by clicking **Information -> Publications**

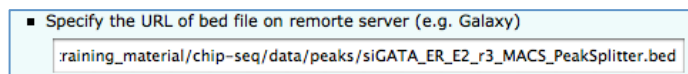
Fetching sequences from UCSC to RSAT

1. In the left frame, click **NGS ChIP-seq -> fetch-sequences from UCSC**
2. Select the Human genome, version hg19



Genome (mandatory) hg19 - Human Feb. 2009 (GRCh37/hg19)

3. Genomic coordinates can be provided in 3 alternative ways:
 - a. Paste the coordinates (not very convenient for large peak sets).
 - b. Specify a URL. This option is generally suitable for importing peaks from a Galaxy server or any other Web site.
 - c. Upload a file from your computer.
4. We will use the second option, since the peak coordinates are available in bed format on the course.
 - a. In a separate window of your browser, open **Practicals -> ChIP-seq -> Datasets -> Peaks for motif analysis**
 - b. Choose one dataset (for example *siGATA_ER_E2_r3_MACS_PeakSplitter.bed*), right-click on a file link and select **"Copy link location"**.
 - c. Come back to the RSAT "fetch sequences" form, and paste the link in the box
- d. Make sure that the
5. Leave all other parameters unchanged and click **"GO"**.



■ Specify the URL of bed file on remote server (e.g. Galaxy)

raining_material/chip-seq/data/peaks/siGATA_ER_E2_r3_MACS_PeakSplitter.bed

Result

The RSAT web server now sends a request to the UCSC genome Browser, to obtain the peak sequences. This should take no more than a few seconds.

The result page displays a table with links:

- your input file (peak coordinates, in bed format);
- the fetched sequences (in fasta format);
- the log file informs you about the parameters (useful for the tractability of your datasets), file locations and processing time.

Tip: store a copy of the fasta sequences and the log file on your computer (you will not need it, but it is always safe to keep a trace).

Discovering motifs in a single ChIP-seq peak set

1. Below the result table of fetch-sequence, click on the link “*peak-motifs*”.
2. The default peak-motifs web form only displays the essential options. There are only two mandatory parameters.
 - a. The title box has been automatically filled with the name of your query file for fetch-sequences.
 - b. The URL is automatically filled with the link to the peak sequences (fasta file from fetch-sequences).
3. We will now modify some of the advanced options in order to fine-tune the analysis according to your data set.
 - a. Open the “*Motif Discovery parameters*” title, and check the oligomer sizes 6 and 7.
 - b. Under “*Compare discovered motifs with databases*”, check the two following items:

☒ JASPAR core Vertebrates”

“JASPAR PBM (UNIPROBE) Mouse”.
 - c. Under “*Locate motifs and export ...*” select

☐ Peak coordinates specified in **fasta headers** of the test sequence file ([Galaxy](#) format)”
 - d. You can indicate your email address in order to receive notification of the task submission and completion. This is particularly useful because the full analysis may take some time.
4. Click “**GO**”. As soon as the query has been launched, you should receive an email indicating confirming the task submission, and providing a link to the future result page.
5. The Web page also displays a link to this page:

The result will become available at
http://rsat.sb-roscoff.fr/tmp/apache/2013/01/17/peak-motifs.2013-01-17.083050_2013-01-17.083050_cpxNrQ/peak-motifs_synthesis.html

6. You can already click on this link. The report will be progressively updated during the processing of the workflow.

Questions

1. Do we discover significant motifs?
2. Are these motifs biologically relevant? In particular, did the program discover motifs related to ER, GATA or Fox-related factors ?

EXERCISES

Differential analysis

1. Fetch the sequences for the PeakSplitter peaks in the two conditions (siNT and siGATA). Save the two sequence files on your computer, since we will need to upload both of them separately for the next step.
 - a. Note: alternatively, you can take note of the two URLs of the fasta files.
2. Run peak-motifs in differential analysis mode using siNT (not treated) as test and siGATA as control. Do you obtain different motifs than with the single peak set analysis?
3. Swap the datasets (siGATA as test, and siNT as control) and redo the analysis. Do you find motifs ? How do they compare with the previous ones?

Negative controls

1. Use the tool “*random sequences*” to generate artificial sequences of the same length and size as the siGATA peaks, and analyse them with peak motifs in single set mode. Do you expect to find significant motifs? Do you obtain significant motifs?
2. Under the title “*Build control sets*”, use the tool “*random genomic fragments*” to select random peaks of the same number and size as the siGATA peak set, and analyse them with peak-motifs in single set mode. Do you expect to find significant motifs? Do you obtain significant motifs?

MEME chips

1. Open a connection to the *MEME software suite*: <http://meme.nbcr.net/meme/>
2. Click on the link to *MEME-chip*.
3. Analyse the same peak sets as above, and compare the discovered motifs with the ones returned by *peak-motifs*.

FURTHER READING

- A general description of the RSAT software suite
 1. Thomas-Chollier, M., Sand, O., Turatsinze, J. V., Janky, R., Defrance, M., Vervisch, E., Brohee, S. & van Helden, J. (2008). RSAT: regulatory sequence analysis tools. *Nucleic Acids Res* 36, W119-27.
 2. Thomas-Chollier, M., Defrance, M., Medina-Rivera, A., Sand, O., Herrmann, C., Thieffry, D. & van Helden, J. (2011). RSAT 2011: regulatory sequence analysis tools. *Nucleic Acids Res* 39, W86-91.
- Description of peak-motif performances, and application to illustrative datasets
 3. Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D. & van Helden, J. (2012). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res* 40, e31.
- A step-by-step protocol to learn using peak-motifs
 4. Thomas-Chollier, M., Darbo, E., Herrmann, C., Defrance, M., Thieffry, D. & van Helden, J. (2012). A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. *Nat Protoc* 7, 1551-68.
- Other protocols for the RSAT suite
 5. Turatsinze, J. V., Thomas-Chollier, M., Defrance, M. & van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat Protoc* 3, 1578-88.
 6. Defrance, M., Janky, R., Sand, O. & van Helden, J. (2008). Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat Protoc* 3, 1589-603.
 7. Sand, O., Thomas-Chollier, M., Vervisch, E. & van Helden, J. (2008). Analyzing multiple data sets by interconnecting RSAT programs via SOAP Web services: an example with ChIP-chip data. *Nat Protoc* 3, 1604-15.