MoCA: Tool for Motif Conservation Analysis

Saket Choudhary, Anton Valouev

valouev@usc.edu

Motifs are short DNA sequences that appear recurrently and often have biological functions. They act as sequence specific binding sites for transcription factors. Motif analysis of ChIP-Seq datasets often reports multiple motifs. However, determining the quality of a reported motif is hard. Motifs reported by motif analysis tools such as MEME[1] can often not be the 'true motifs' and can have significant p-value(or E-values) for even 'false motifs'.

Another metric often used to filter out motifs involves calculating the distance of the ChIP-seq peak with the center of the reported motif. This involves reporting the motifs appearing in ± 100 base pairs of the peak. However, with this approach will often not work for identifying new co-transcription factor motifs.

Any metric to assess the quality of motifs, should also rely on biological relevance besides the statistical analysis. Since the motif acts as a specific binding sequence, it can be expected to be conserved evolutionarily. We hence hypothesised that, a 'true motif' should exhibit high Phylop[3] and Gerp[2] conservation scores. In order to test the hypothesis, we developed MoCA, a tool to perform conservation analysis of reported motifs. MoCA makes use of the Phylop and Gerp scores to assess the conservation profile of the motif bases and compares it with flanking bases and by searching for motifs in random genomic regions. If our hypothesis is true, the motif bases should show significantly more conservation as compared to the bases flanking the motifs on either side.

We performed analysis on various ENCODE Chip-Seq datasets and found that the 'true motifs', validated experimentally do exhibit high conservation scores. A summary of the workflow and results for GATA1 are elaborated in Figure 1.

MoCA is available as a web service at `http://moca.usc.edu` MoCA has an inbuilt support for directly analysing ENCODE Chip-Seq datasets, where the conservation analysis plots an be generated by specifying the ENCODE experiment id.

References

[1] T. L. Bailey, J. Johnson, C. E. Grant, and W. S. Noble. The meme suite. *Nucleic acids research*, page gkv416, 2015.

[2] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLoS Comput Biol*, 6(12):e1001025, 2010.

[3] A. Siepel, K. S. Pollard, and D. Haussler. New methods for detecting lineage-specific selection. In *Research in Computational Molecular Biology*, pages 190–205. Springer, 2006.