# MoCA: Tool for motif conservation analysis

Anton Valouev, Saket Choudhary

valouev@usc.edu

Motifs are short DNA sequences that appear recurrently and often have biological function such as acting 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. Measures such as p-value(or E-values) reported by motif analysis tools such as MEME[cite] can often be significant for even false motifs. (How do we defined 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.

In order to come with a metric to assess the quality of motifs, which also takes into account the biological function. 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[**?**, ]nd Gerp[**?**, ]onservation 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.