ENCODE motif enrichment antibody characterization method – February 2016

The motif enrichment analysis was done by Zhizhuo Zhang (zhizhuo@mit.edu) using a custom motif enrichment pipeline (https://github.com/zzz2010/ENCODE_motifvalidation, version FEB-24-2016) based upon a collection of known motifs available at http://compbio.mit.edu/encode-motifs/ and previously published TF ChIP-seq peak data (http://www.broadinstitute.org/~zzhang/motifpipeline/data/TrainSetInfo.txt) retrieved from the Cistrome Database (http://cistrome.org/db/#/). Three types of motif enrichment scores were computed by overlapping the motif instances with the given ChIP-seq peak locations, which included a global enrichment z-score (compare the actual motif with the shuffled-version motif), a positional bias z-score (compare peak center to the peak flanking region, +/-100bp), and a peak rank bias z-score (compare the high signal value regions to the low signal value regions). Then, a combined enrichment score was derived by taking the average of the three enrichment scores listed above. Next, the known motifs were grouped as 282 clusters by their PWM similarities and each motif cluster was ranked by the highest combined enrichment score of that motif cluster. Hence, each known motif was assigned to the ranking of the motif cluster it belonged to.

The method is based on a Bayesian approach developed by Zhang, Kheradpour et al. that assumes the enrichment ranking distribution of the known motif follows a mixture of two negative-binomial distributions, corresponding to two cases: 1) the antibody used in the ChIP-seq experiment from which the input peak calls were derived targets the TF corresponding to that motif, and 2) the antibody used in the ChIP-seq experiment is actually targeting other TFs different from the query. If the two negative-binomial distributions for the tested motif are known, then we can derive the accept/reject probability of the tested antibody given the enrichment ranking of that motif.

The parameters of two negative-binomial distributions are estimated from a collection of previously published ChIP-seq data available for the query TF and all other TFs (stored in pipeline_script/cistrome_model.pickle commit SHA: 9ff2e851f1fb97ae3bfc4287c3af554583eca994). This accept probability calculation also assumes that the prior probabilities of passing validation is 50%, and takes into account that different TFs may share the same motif and one TF may use multiple motifs.

The current ENCODE antibody characterization standards uses “accept probability” > 0.6 as criteria for secondary antibody characterization.