Are computationally predicted footprints result of DNase I cleavage bias?

Eduardo G. Gusmao, Martin Zenke and Ivan G. Costa
Institute for Biomedical Engineering, RWTH Aachen University Medical School, Aachen, Germany

DNase I digestion followed by massive sequencing (DNase-seq) has proven to be a powerful technique for identifying active transcription factor (TF) binding sites on a genome-wide scale [1, 2, 3, 4]. Several computational approaches have been proposed to find nucleotide-resolution footprints, regions with 5 to 20 bps within two DNase-seq peaks [2, 4]. Recently, He et al. (2014) demonstrated that DNase-seq signal has biases reflecting the preference of DNase I to cleave particular sequences. Moreover, they show that the performance of a digital footprint method correlates with the cleavage bias of the underlying TF motif and that footprints are outperformed by simple DNase hypersensitivity sites tag count scoring (DHS-TC). However, these results were based on footprints predicted with a simple version of the digital footprint occupancy score (FOS) from [4] and no attempt was made to correct sequence bias previous to footprint prediction.

To address these questions, we extended our segmentation-based digital footprinting framework (HINT – HMM-based identification of TF footprints) [2] by performing bias correction of DNase-seq signals (HINT-BC). We estimated DNase I cleavage bias as in [3] on ENCODE DNase-seq data sets obtained from Crawford lab (H1-hESC, HeLa-S3, Huvec and K562) and Stamatoyannopoulous lab (HepG2, Huvec and K562). We observed that cleavage bias is distinct for each DNase-seq data set and that differences were larger between experiments from distinct labs. We then executed HINT, HINT-BC, DHS-TC and FOS on these data sets and evaluated predictions with 139 TF ChIP-seq data sets measured on these cell types. Performance of methods were evaluated regarding their area under the ROC curve (AUC) at 10% false positive rate. Results indicate that HINT-BC significantly outperforms all compared methods, while FOS was outperformed by all methods (Friedman-Nemenyi hypothesis test at 0.05 significance level). This reinforces our point that the method evaluated in [3] is not a good representative of footprint detection methods and that footprint methods profit from sequence bias correction.

Next, we measured the correlation between observed and expected number of DNase cleavage sites around each TF. This statistics measures the potential "bias score" of a TF motif for a given DNase-seq assay [3] (Fig. 6). We observed a high negative correlation between FOS AUC and the "bias score" (0.41, p-value $< 10^{-5}$) for all evaluated motifs, which agrees with the observation that FOS footprints are affected by DNase cleavage bias. HINT and HINT-BC presented negative correlation values of -0.14 and -0.04 (p-values > 0.05). These results show that the impact of DNase-seq cleavage bias is low on robust digital footprinting methods and can be further decreased after the correction of DNase-seq signal.

[1] Crawford, G.E. *et al.* (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, **16**(1), 123–131.

[2] Gusmao, E.G. *et al.* (2014). Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics*, pages btu519+.

[3] He, H.H. *et al.* (2014). Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat Meth*, **11**(1), 73–78.

[4] Neph, S. *et al.* (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**(7414), 83–90.