Prediction of genome-wide in vivo transcription factor binding using factor-specific DNase footprinting models

Galip Gürkan Yardımcı[1], Gregory E. Crawford[1,2], Uwe Ohler[1,3]

Keywords: gene regulation, DNase-seq, footprinting, transcription factor binding

The identification of DNase I hypersensitive sites and DNase footprints are well established methods for identification of genomic regulatory regions and DNA-protein interactions, respectively. Using data generated by high throughput DNase-seq assays, we propose models to identify binding locations of transcription factors in different cell lines in a genome-wide manner by modeling each factor's unique DNase footprint. Contrary to most existing approaches, our model aims to represent the footprint shape in detail while trying to account for the contribution of overall DNase hypersensitivity around a binding site to assess the accuracy of the footprints by themselves – a necessary feature to identify specific sites bound under different conditions. We model each transcription factor's footprint using two features: distribution of DNase-seq reads at each base and the DNase-seq coverage. Transcription factor binding predictions are validated rigorously using ChIP-seq assays from the ENCODE consortium. We achieve a mean AUC value of 95% for 20 transcription factors. We find that AUC values tend to depend on quality of motif associated with transcription factor and transcription factor structural family. For each transcription factor, we show that some ChIP-seq peaks do not overlap with a DNase footprint and characterize such peaks according to ChIP-seq signal intensity and co-binding proteins.

[1] Institute for Genome Sciences & Policy, Duke University, Durham, NC

[2] Department of Pediatrics, Duke University, Durham, NC

[3] Department of Biostatistics & Bioinformatics, Duke University, Durham, NC