**Are Computationally Predicted Footprints Result of DNase I Cleavage Bias?**

**Eduardo G. Gusmão**\***, Martin Zenke and Ivan G. Costa**

***Institute for Biomedical Engineering, RWTH Aachen University Medical School, Aachen, Germany***

DNase I cleavage followed by massive sequencing (DNase-seq) has proven to be a powerful genome-wide technique for identifying active transcription factor (TF) binding sites [1–4]. Several computational approaches have been proposed to find nucleotide-resolution footprints (5-20 bp regions within two DNase-seq peaks) [3–7]. 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 (footprint occupancy score – FOS) [3] correlates with the cleavage bias of the underlying TF motif and that footprints are outperformed by simple DNase-seq tag count scoring (TC). Here, we test these claims using more sophisticated digital genomic footprinting methods. Furthermore, we verify if it is possible to improve computational methods by correcting such DNase I cleavage bias.

Estimation of intrinsic DNase I cleavage bias was performed by calculating ratios of observed cleavage sites to background k-mers in DNase-seq hypersensitivity sites (DHSs) [2].

Fig. 1: Correlation of bias scores between different DNase-seq datasets given all possible DNA 6-mers.

DNase I Cleavage bias correction was based on smoothed versions of both DNase I and bias score signals [2].

Fig. 2: Observed cleavage, bias score and corrected signal for TFs E2F4 and EGR1. Signals were standardized to be in [0,1].

Fig. 3: (A,B) Performance of methods as ROC curves for TFs E2F4 and EGR1. In the legend it is shown the AUC at 10% FPR. (C) Friedman-Nemenyi hypothesis test. Each row starts with the Friedman ranking for each method. A shadowed cell means that the method in the column outperformed the method in the row (95% confidence level). (D) Correlation between the performance of each method (in relation to the DNase I TC) and the OBS (correlation between observed and bias signal).

We applied the digital footprinting method HINT [4] to the DNase-seq signal and the bias corrected (BC) signal. We observed that bias corrected version of HINT – HINT (BC) – outperformed all other methods, including the site-centric tag count (TC), footprint occupancy score (FOS), position weight matrix (PWM) bitscore, Boyle [5], Neph [3], Cuellar [6] and Centipede [7] (Fig.3C). Interestingly, the Friedman-Nemenyi test also showed that the bias corrected (BC) version of HINT significantly outperforms the original version. We also evaluated the correlation between the performance of each method (represented by their AUC relative to TC’s AUC) and the observed vs bias signal (OBS) (Fig.3D) [2]. The latter corresponds to the correlation between the observed DNase I cleavage and bias score (Fig.2). Significant negative correlations were observed for FOS, Boyle and Neph. Again, since HINT (BC) portrayed a smaller correlation than HINT, the bias correction demonstrated to mitigate prediction biases.