Correction of DNase-seq cleavage bias impacts on quality of footprinting

Eduardo G. Gusmao1,2, Martin Zenke1,2, Ivan G. Costa1,2,3,\*.

1 IZKF Computational Biology Research Group, RWTH Aachen University Medical School, Aachen, Germany.

2 Department of Cell Biology, Institute of Biomedical Engineering, RWTH Aachen University Medical School, Aachen, Germany.

3 Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Germany.

\* e-mail: [ivan.costa@rwth-aachen.de](mailto:ivan.costa@rwth-aachen.de)

**To the editor:** Recently, He et al.1 published an important study that revisits current DNase-seq protocols. They demonstrate for the first time that intrinsic DNase-seq cleavage bias around transcription factor binding sites (TFBSs) affects the performance of computational footprinting. They show that the accuracy of a footprinting method (area under the ROC curve, AUC) inversely correlates with the amount of DNase-seq cleavage bias in 36 transcription factors (TFs). Furthermore, the authors claim that counting the number of DNase-seq reads around putative TFBSs (tag count), which is the simplest method for detection of active binding sites possible, outperforms computational footprinting. Such claim relies on the evaluation of a simplistic footprinting method: the footprint score (FS)1. Moreover, no bias correction was performed prior to computational footprinting.

To evaluate the influence of cleavage bias on state of art footprinting methods, we reproduced and extended the analysis presented in Fig. 6b of He et al.1. We included six additional footprinting methods in our analysis: HINT3, Boyle4, Neph2 Centipede5, Cuellar6 and PWM bit-score. We also applied HINT on a bias-corrected DNase-seq signal (HINT bias-corrected, HINT-BC). The bias correction followed the TF-centric 6-mer scheme presented in He et al.1. We evaluated these methods on a benchmarking data set based on 83 TFs on H1-hESC and K562 cells3 (Supplementary Notes).

The Boyle, Neph, FS and PWM methods presented a significant negative correlation (R=-0.31, R=-0.3, R=-0.22 and R=-0.2, respectively) between their accuracy performance and amount of DNase-seq cleavage bias (Fig. 1a; p-value<0.05). Moreover, HINT-BC, which is the only method performing bias correction, performed best and displayed the lowest absolute correlation (R=-0.06). Concerning prediction accuracy, several footprinting methods (Boyle, HINT, HINT-BC and Neph) have a higher AUC than tag count (Fig. 1a and supplementary Fig.4; p-value<0.05). We also observed that HINT-BC significantly outperformed all other methods (p-value<0.05). Altogether, our results contradict He et al., as several footprinting methods are either not significantly influenced by cleavage bias or superior than the tag count method (Fig. 1a).

As an example, we show corrected and uncorrected DNase-seq average profiles around TFBSs with highest AUC gain between HINT-BC and HINT (Fig. 1b and c). The NRF1 and EGR1 DNase-seq profiles indicate that the bias-corrected signal fits better the TFs DNA affinity sequence than the uncorrected signal (Supplementary Fig. 5).

The refined DNase-seq protocol and DNase cleavage bias presented in He et al. are of great importance to the regulatory genomics field. However, robust *in silico* techniques are also crucial to correct for experimental artifacts and derive valid biological predictions. We demonstrate that four footprint methods are more accurate than a simple read counting approach. Furthermore, the correction of DNase-seq signal virtually removes the effects of the cleavage bias on computational footprinting.

**Supplementary Information:** Please find attached supplementary information regarding computational experiments and further results.

**Competing Financial Interests:** The authors declare no competing financial interests.

1 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.

2 Neph, S. et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature, 489(7414), 83-90.

3 Gusmao, E.G. et al. (2014). Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. Bioinformatics, 30(22), 3143-3151.

4 Boyle, A.P. et al. (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Research, 21(3), 456-464.

5 Pique-Regi, R. et al. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research, 21(3), 447-455.

6 Cuellar-Partida, G. et al. (2012). Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics, 28(1), 56-62.

**Figure 1** Performance of footprinting methods and examples of DNase-seq profiles. (**a**) Correlation between the performance of eight footprinting methods and their bias estimated for 83 transcription factors of the cell types H1-hESC and K562. The x-axis represents the correlation between the uncorrected and bias signal; higher values indicate higher bias. The y-axis represents the ratio between the AUC at 10% false positive rate for each method and the tag count method; higher values indicate higher accuracy. (**b**-**c**) Average uncorrected and bias-corrected DNase-seq signal around motifs enriched with ChIP-seq for the transcription factors NRF1 (**b**) and EGR1 (**c**). Signals were standardized to be in [0,1]. The motif logo represents underlying DNA sequences centered on the TFBSs.