Addressing DNase-seq cleavage bias and residence time on computational footprinting

Eduardo G. Gusmao[1,2], Martin Zenke[1,2], Ivan G. Costa[1,2,3,*].

[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

Word Count: Abstract – 77 of 70; Document – 1398 of 1500 (including abstract)

Abstract: Experimental artifacts as DNase cleavage bias and transcription factor residence time impact computational analyses of DNase-seq experiments. We investigated these artifacts in a comprehensive panel of DNase-seq data sets employing 14 footprinting methods. We demonstrate that correcting the DNase-seq signal based on cleavage bias estimation significantly improves accuracy of computational footprinting. We also propose a score to detect footprints arising from transcription factors with low residence time, as footprints of such factors have a low predictive performance.

Next-generation sequencing (NGS) combined with genome-wide mapping techniques, such as DNase-seq, contributed greatly to our understanding of gene regulation and chromatin dynamics[1]. DNase-seq allows the nucleotide-level identification of transcription factor binding sites (TFBS) based on the computational search of footprint-like DNase I cleavage patterns on the DNA[1-3]. A number of computational methods have been proposed for this task[4-12]. They can be categorized in (1) segmentation-based methods[4-8] (the genome is scanned for footprint predictions based solely on DNase-seq data) and (2) site-centric methods[9-12] (DNase-seq data around motif-predicted binding sites [MPBSs] are obtained and used to classify these sequence-based predictions as being bound or unbound).

NGS-based data are significantly affected by biases, which are inherent to the experimental protocols used[13,14]. One major artifact of DNase-seq experiments is the cleavage bias, which is due to DNase I having different binding affinities towards specific DNA sequences[15]. Recently, He et al.[13] showed that intrinsic DNase I cleavage bias around TFBSs strongly affects the performance of a computational footprinting method (footprint score; FS) on a factor-specific manner. 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. They also indicate several transcription factors (TFs), such as nuclear receptors, where the DNase-seq profile resembles their cleavage bias estimate. Furthermore, they indicate that counting the number of DNase-seq reads around putative TFBSs (tag count; TC) outperforms the evaluated computational footprinting method.

Another experimental aspect affecting the computational analysis of DNase-seq is the residence time of TF binding. Sung et al.[7] showed that short-lived TFs display a lower DNase I cleavage protection pattern (low number of DNase reads surrounding the footprint). Moreover, they also noticed that short-lived TFs as nuclear receptors have DNase-seq profiles resembling cleavage bias estimates. While both He et al.[13] and Sung et al.[7] show the challenges imposed by cleavage bias and residence time in the computational analysis of DNase-seq data, there is so far no comprehensive study showing their impact in state-of-the-art footprinting methods. Moreover, with the exception of a few methods accounting for cleavage bias[7,12,15], there has been no attempts to address these issues computationally.

To evaluate the influence of cleavage bias and residence time on state-of-the-art footprinting methods, we reproduced and extended the analysis by He et al.[13]. We included ten additional footprinting methods: Neph[4], Boyle[5], Wellington[6], DNase2TF[7], HINT[8], Centipede[9], Cuellar[10], PIQ[11], FLR[12] and PWM bit-score. We also applied HINT on bias-corrected DNase-seq signals. Such correction was performed based on the estimation of cleavage bias within DNase I hypersensitivity sites (HINT bias-corrected; HINT-BC) or deproteinized DNase-seq experiments (HINT bias-corrected on naked DNase; HINT-BCN). The bias correction followed an adaptation of the TF-centric 6-mer scheme presented in He et al.[13]. We evaluated these methods on a benchmarking data set based on 88 TFs on H1-hESC and K562 cells[8] (online methods).

Our analysis shows that only six out of 13 evaluated methods (Wellington, Neph, Boyle, DNase2TF, Centipede and FS) present a significant negative Pearson correlation (R = -0.35, R = -0.33, R = -0.28, R = -0.28, R = -0.24 and R = -0.22, respectively) between their accuracy performance and amount of DNase-seq cleavage bias (Fig. 1a; adjusted p-value < 0.05). Equivalent results are also observed on the same TFs and cellular conditions analyzed in He et al. (Supplementary Fig. 1). As expected, methods explicitly using 6-mer cleavage bias statistics (HINT-BC, HINT-BCN and FLR) or performing smoothing (PIQ, Cuellar) are not significantly influenced by cleavage bias (Supplementary Table 2). Moreover, HINT-BC displayed the lowest absolute correlation over all methods (R = -0.06). As an example, we show bias estimates, corrected and uncorrected DNase-seq average profiles around TFBSs with highest AUC gain between HINT-BC and HINT (Fig. 1b and c; Supplementary Fig. 2). The NRF1 and EGR1 DNase-seq profiles indicate that the bias-corrected signal fits better their sequence affinity than the uncorrected signal. These TFs have cleavage bias estimates with an inverse footprint profile and motifs rich on G/A nucleotides.

Ivan Gestaira Costa …, 7/23/15 11:43 AM

Comment: We should maybe include here a few words about the cleveage estimates thenselves. They should be CA poor or? Maybe we need to report of these tables as well.

To further explore the nature of the cleavage bias, we performed a clustering on cleavage bias estimates of 15 DNase-seq datasets (Supplementary Table 1), which include 3 deproteinized DNA experiments (Supplementary Fig. 3). It is possible to discriminate the cleavage bias estimated on DNase I hypersensitivity sites (DHS) by protocol (single hit[2] *vs.* double hit[3]). Naked DNase-seq experiments are grouped together and have a moderate correlation with DHS-estimated cleavage bias from corresponding protocols.

Concerning prediction accuracy, all segmentation-based footprinting methods (HINT-BC, HINT-BCN, HINT, Boyle, DNase2TF and Wellington) and one site centric method (PIQ) have a higher AUC at 10% false positive rate than TC (Supplementary Fig. 4; Supplementary Table 2 and 3; adjusted p-value < 0.05; Friedman-Nemenyi test). We also observed that HINT-BC has highest AUC values and significantly outperforms all methods with the exception of HINT-BCN (adjusted p-value < 0.05). Altogether, our results demonstrate that several footprinting methods outperform simple TC approach and that signal correction based on DHS-estimated cleavage bias is the best approach to mitigate the impact of intrinsic DNase-seq cleavage bias.

Concerning TFs with low residence time, as nuclear receptors, we observe that they have poor DNase-seq footprint profiles and this can only be partially addressed by bias correction (Supplementary Fig. 4). Indeed, evaluated footprinting methods have low AUC values for these factors (bottom quartile AUC values for HINT-BC and TC; Supplementary File 1). To investigate the effects of TF residence time on footprint predictions, we propose a statistic inspired on the concepts presented in Sung et al.[7]. The protection score measures the difference between the amounts of DNase I digestion in the flanking regions of TFBSs and within the binding site on cleavage bias corrected DNase signals (online methods). Therefore, we can analyze the performance of methods on TFs with distinct residence times. For this, we used an expanded data set with 233 combinations of DNase-seq experiments and TFs (online methods).

We observed that TFs with known short residence time on DNA, such as nuclear receptors AR, ER and GR, present a negative protection score (Fig. 2a). TFs with intermediate and long

residence time on DNA (C-jun and CTCF, respectively)[7] present a positive protection score. The amount of protection is clearly reflected in the bias corrected DNase-seq profiles (Fig. 2b, 2c and 2d). In addition, Fig. 2a also reveals an association of the protection score and the AUC of HINT-BC. Overall, the protection score positively correlates (Spearman correlation) with the AUC values of evaluated methods, such as TC (R = 0.19) and HINT-BC (R = 0.26), and negatively correlates (R = -0.49) with the DNase-seq cleavage bias (adjusted p-value < 0.05). These results reinforce the concept that short residence time indeed imposes a challenge to footprinting methods. Clearly, patterns of raw DNase-seq signals around such factors are likely to be artifact of cleavage bias.

The refined DNase-seq protocol and DNase cleavage bias presented in He et al.[13] and TF binding time presented in Sung et al.[7] underscore that robust *in silico* techniques are required to correct for experimental artifacts and to derive valid biological predictions. We demonstrate that seven footprinting methods are more accurate than a simple read counting approach. Furthermore, the correction of DNase-seq signal or smoothing virtually removes the effects of the cleavage bias on computational footprinting. While it is a challenge to predict footprints of TFs with short residence time, we show that the protection score can identify predictions of such TFs in order to improve the confidence assessment of footprint predictions.

Methods and Supplementary Information: Supplementary information regarding methods, computational experiments and further results are found in attachment. Software, scripts and benchmarking data are available in www.costalab.org/hint-bc.

Acknowledgements: This work was supported by the Interdisciplinary Center for Clinical Research (IZKF Aachen), RWTH Aachen University Medical School, Aachen, Germany.

Author Contributions: E.G., M.Z. and I.C. designed the research. E.G. wrote HINT program code. E.G. and I.C. analyzed data. E.G., M.Z. and I.C. wrote the manuscript.

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

[1] ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489(7414), 57-74.
[2] 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.
[3] Sabo, P.J. et al. (2004). Genome-wide identification of DNase I hypersensitive sites using active chromatin sequence libraries. PNAS, 101(13), 4537-4542.
[4] Neph, S. et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. Nature, 489(7414), 83-90.
[5] 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.
[6] Piper, J. et al. (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Research, 41(21), e201.
[7] Sung, M.-H.H. et al. (2014). DNase footprint signatures are dictated by factor dynamics and DNA sequence. Molecular Cell, 56(2), 275-285.
[8] 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.
[9] 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.
[10] Cuellar-Partida, G. et al. (2012). Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics, 28(1), 56-62.
[11] Sherwood, R.I. et al. (2014). Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nature Biotechnology, 32(2), 171-178.
[12] Yardımcı, G.G. et al. (2014). Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Research, 42(19), 11865-11878.
[13] 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.
[14] Meyer, C. and Liu, X. (2014). Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nature Reviews Genetics, 15(11), 709-721.
[15] Hesselberth, J.R. et al. (2009). Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. Nature Methods, 6(4), 283-289.

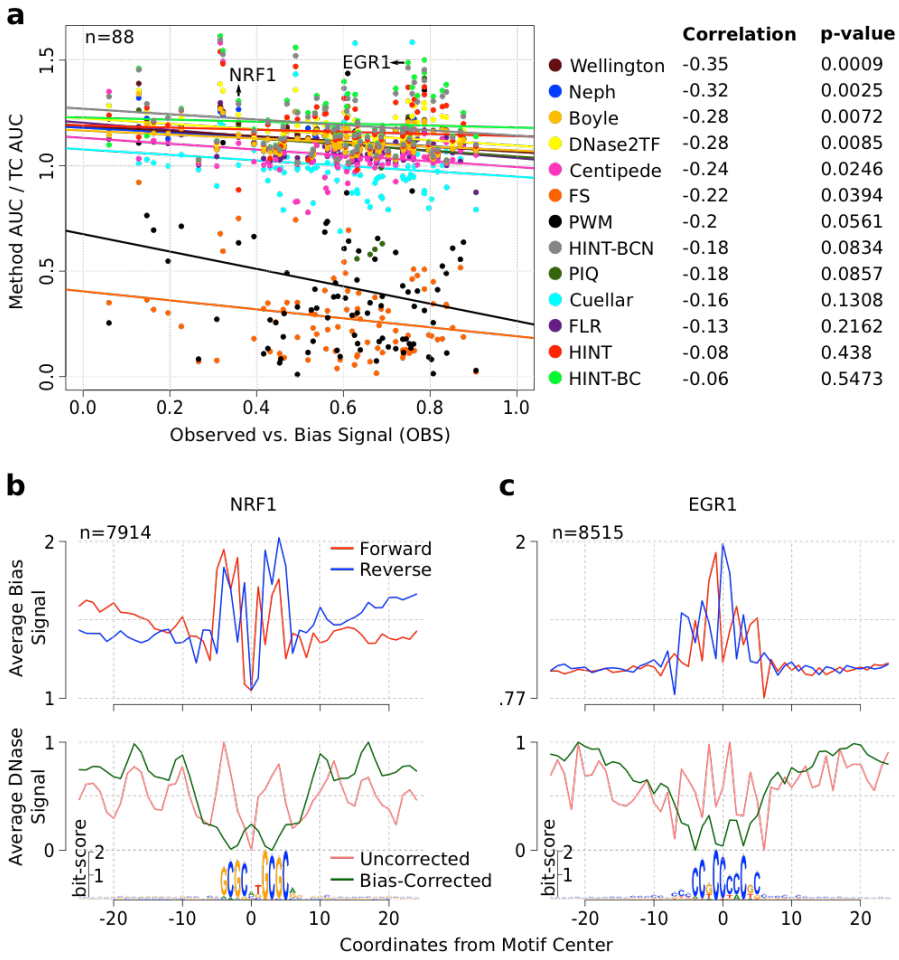

Figure 1 Performance of footprinting methods and examples of DNase-seq profiles. (**a**) Pearson correlation between the performance of 13 footprinting methods and their cleavage bias estimated for 88 transcription factors of the cell types H1-hESC and K562. The x-axis represents the correlation between 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 (TC) method; higher values indicate higher accuracy. (**b**-**c**) Average bias signal (top) and uncorrected/corrected DNase-seq signal (bottom) for the transcription factors NRF1 (**b**) and EGR1 (**c**). Signals in the bottom graph were standardized to be in [0,1]. The motif logo represents underlying DNA sequences centered on the TFBSs.

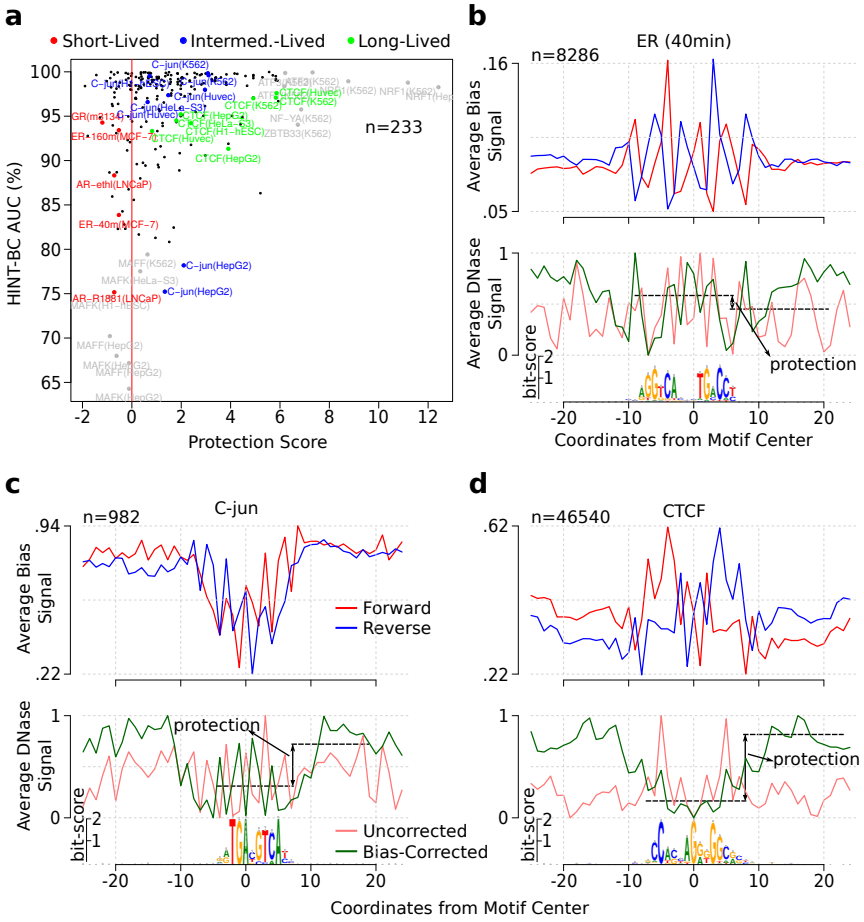

Figure 2 Impact of transcription factor residence binding time on computational footprinting. (**a**) Scatter plot with the protection score (x-axis) vs. AUC of HINT-BC (y-axis) for 233 TFs binding on 11 cell types. We indicate in red experiments with nuclear receptors AR, ER and GR (short residence time); in blue experiments with C-jun (intermediate residence time); in green experiments with CTCF (long residence time) and in gray experiments with either high protection score (> 6) or low AUC values (< 0.8). (**b-d**) Average bias signal (top) and uncorrected/corrected DNase-seq signal (bottom) for the transcription factors ER (**b**), C-jun (**c**) and CTCF (**d**). Signals in the bottom graph were standardized to be in [0,1]. The motif logo represents underlying DNA sequences centered on the TFBSs.