HINT-BC – HMM-based Identification of Transcription Factor Footprints on Bias-Corrected DNase-seq Data

Eduardo G. Gusmao [1,2,†], Martin Zenke [2] and 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.

[†] Presenting author [*] ivan.costa@rwth-aachen.de

1 Introduction

Advances on next-generation sequencing techniques have enabled investigators to develop high-throughput versions of many essays such as chromatin immunoprecipitation (ChIP-seq) and DNase I footprinting (DNase-seq). The latter, a technique in which digested chromatin fragments are sequenced to identify accessible regulatory regions, can be used to accurately identify transcription factor (TF) binding sites through a well-characterized pattern of DNase I cleavage [1–4]. In general, transcription factor binding sites (TFBSs) present high levels of DNase I cleavage activity at their flaking regions and low cleavage activity at the positions where proteins are bound (footprints) [5]. Recently, many computational footprinting methods have been developed to automatically detect TFBSs based on DNase-seq data [3, 5–8].

Nevertheless, it is known that DNase-seq exhibits an intrinsic DNase I cleavage bias [9, 10]. Such bias reflects the binding preference of the DNase I enzyme to certain *k*-mers. Since TFs present binding affinity towards certain DNA sequences, footprints are likely to be affected by DNase-seq cleavage bias. Previous attempts to correct such bias did not result in significant performance improvement [2].

Here, we couple a DNase-seq cleavage bias correction strategy with our previous approach, termed HINT (HMM-based identification of transcription factor footprints) [8]. We compare our novel method – HINT bias-corrected (HINT-BC) – with seven recent computational footprinting methods in the literature and with its uncorrected version. We study whether bias was mitigated, the impact on computational footprinting performance and the nucleotide-level changes on DNase-seq profile shape for many transcription factors.

2 Method

We performed a modified *k*-mer-based bias-correction strategy as presented in [9]. Briefly, the DNase-seq cleavage bias is estimated from the aligned reads inside DNase hypersensitivity sites (DHSs). The observed cleavage score for a particular *k*-mer *w* is defined as the number of cleavages that occurred centered on *w*. The background cleavage score equals the number of times *w* occurs. The bias estimation for *w* corresponds to the ratio between the observed and background cleavage scores. Finally, the bias-corrected DNase-seq signal was created based on the original DNase-seq signal and smoothed versions of the DNase-seq signal and the estimated bias signal.

In this analysis we considered six competing methods: Boyle [5], Neph [3], Centipede [6], Cuellar [7], ranking of TFBSs based on the total number of surrounding DNase-seq reads (referred to as tag-count; TC) and footprint score (FS) [9]. Furthermore, we used the position weight ma-

trix (PWM) motif matching bit-score as a 'control' scoring. We compared these methods with HINT [8] and the HINT bias corrected (HINT-BC) version. This comprehensive comparison totals nine methods.

We used DNase-seq aligned reads from ENCODE [4] (Crawford lab) cell types H1-hESC and K562. To create the validation data set we obtained ChIP-seq enriched regions (peaks) from ENCODE Analysis Working Group (AWG) track and PWMs from Jaspar, Uniprobe and Transfac repositories. The validation data set consists of all putative binding sites obtained by matching the PWMs on the genome. These putative binding sites are considered true TFBSs if they have ChIP-seq evidence and false TFBSs otherwise.

3 Results

First we compared the amount of bias with the performance of the methods for K562 cell type. For that, we evaluated the observed *v*s. bias signal (OBS), which can be defined as the Pearson correlation between the average profile of these two signals for all putative binding sites with ChIP-seq evidence. Then, we evaluated the Pearson correlation between the OBS and the area under the ROC curve (AUC) for each TF. We observed that only four out of eight methods (FS, PWM, Boyle and Neph) presented a significant (p-value < 0.05) negative correlation (-0.17, -0.22, -0.3 and -0.26, respectively), i.e. the performance of these methods are being negatively influenced by the DNase-seq cleavage bias. Centipede, Cuellar, HINT and HINT-BC presented a Pearson correlation of -0.12, -0.1, -0.05 and -0.03. This is evidence that some computational footprinting methods implicitly correct the bias through signal smoothing. Moreover, the HINT-BC method presented the lowest absolute correlation among all methods, including the non-corrected HINT. This suggests that the bias-correction strategy mitigates the DNase-seq cleavage bias.

Next, we evaluated the performance of the methods with regard to their AUC at the 10% false positive rate (Fig. 1A and B). We applied the Friedman-Nemenyi test to compare the AUC values of distinct methods. We observed that methods which are based on segmenting the genome with window-based statistics or HMMs (Boyle, Neph and HINT) outperformed (p-value < 0.05) methods that classify motif-predicted putative binding sites as bound or unbound (TC, FS, Cuellar, Centipede and PWM) in the specificity level. Also, HINT-BC outperformed all other methods (p-value < 0.05). Such result demonstrates that a bias-correction strategy improves digital genomic footprinting. However, we point to the fact that some complex approaches such as Centipede and Cuellar are outperformed by the TC method, which does account for nucleotide-level footprint patterns.

The bias correction led to substantial change in the average DNase I cleavage patterns surrounding the TFs. The Fig. 1C shows examples of such changes for selected TFs in cell type K562. We observed that the bias-corrected DNase-seq signal fits the high affinity regions of the TF motifs. In contrast, uncorrected DNase-seq signal presents a higher signal in the center of the motif, which does not reflect the affinity regions so clearly. Such patterns reflect bias corrections which are beneficial to footprinting method accuracy.

Recently, the performance of computational footprinting techniques was questioned and shown to be outperformed by simple statistics such as ranking putative binding sites based on the number of DNase-seq tag counts within its vicinity [9]. However, we show that our approach successfully address the intrinsic DNase-seq cleavage bias and outperforms a number of competing methods in the literature.

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

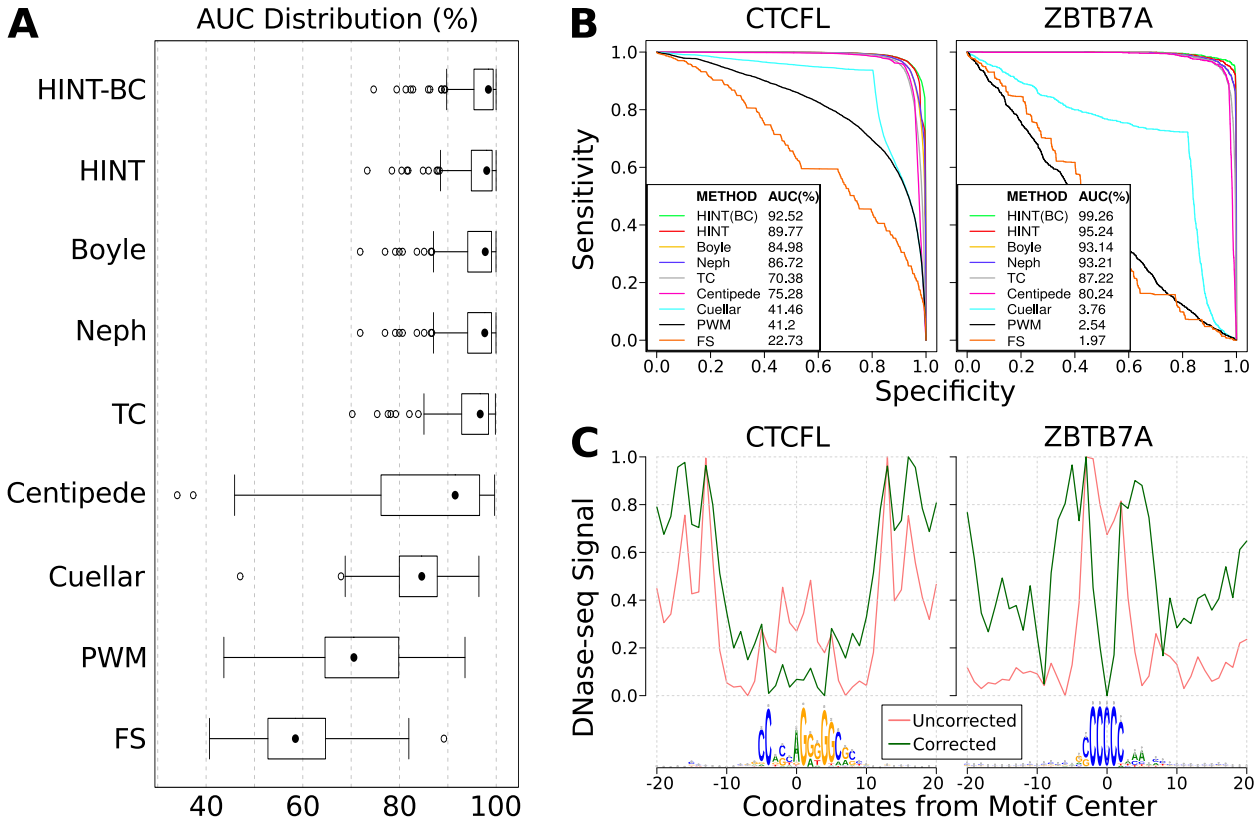

Fig. 1: (A) Distribution of AUCs for all methods tested. (B) ROC curves for selected TFs. The AUC at 10% FPR is shown for all methods. (C) Average uncorrected and bias-corrected DNase-seq signals around selected TFs with ChIP-seq evidence. On the bottom, it is shown the motif logo for all DNA fragments within these regions.

References

[1] Crawford, G. E. *et al.*, "Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)," *Genome Research*, vol. 16, no. 1, pp. 123–131, Jan. 2006.

[2] Hesselberth, J. R. *et al.*, "Global mapping of protein-DNA interactions in vivo by digital genomic footprinting," *Nature Methods*, vol. 6, no. 4, pp. 283–289, Mar. 2009.

[3] Neph, S. *et al.*, "An expansive human regulatory lexicon encoded in transcription factor footprints," *Nature*, vol. 489, no. 7414, pp. 83–90, Sep. 2012.

[4] ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome." *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012.

[5] Boyle, A. P. *et al.*, "High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells," *Genome Research*, vol. 21, no. 3, pp. 456–464, Mar. 2011.

[6] Pique-Regi, R. *et al.*, "Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data," *Genome Research*, vol. 21, no. 3, pp. 447–455, Mar. 2011.

[7] Cuellar-Partida, G. *et al.*, "Epigenetic priors for identifying active transcription factor binding sites," *Bioinformatics*, vol. 28, no. 1, pp. 56–62, Jan. 2012.

[8] Gusmao, E.G. *et al.*, "Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications," *Bioinformatics*, vol. 30, no. 22, pp. 3143–3151, Nov. 2014.

[9] He, H.H. *et al.*, "Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification," *Nat Meth*, vol. 11, no. 1, pp. 73–78, Jan. 2014.

[10] C. Meyer and X. Liu, "Identifying and mitigating bias in next-generation sequencing methods for chromatin biology." *Nature Reviews. Genetics*, 2014.