Analysis of computational footprinting methods for DNase sequencing experiments

Eduardo G. Gusmao[1,2], Manuel Allhoff[1,3], 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

Editorial summary

The comparison of 14 computational methods for detecting transcription factor binding sites in DNase hypersensitive regions in the genome determines which methods work consistently well, how cleavage bias should be corrected for and what is the best score to rank methods.

Abstract: DNase-seq allows a nucleotide-level identification of transcription factor binding sites based on the computational search of footprint-like DNase I cleavage patterns on the DNA. Frequently, in high-throughput methods, experimental artifacts like DNase I cleavage bias impact ~~on~~ the computational analysis of DNase-seq experiments. Here we performed a comprehensive and systematic study on the performance of computational footprinting methods. We evaluated 14 footprinting methods on a panel of DNase-seq experiments for their ability to recover cell-specific transcription factor binding sites. We show that three methods: HINT, DNase2TF and PIQ consistently outperform other evaluated methods. We demonstrate that correcting the DNase-seq signal for experimental artifacts significantly improves accuracy of computational footprints. ~~Moreover, ranking footprints by the number of reads in their vicinity is the best strategy.~~ We also propose a score to detect footprints arising from transcription factors with potentially short residence time, ~~as DNase-seq has poor predictive performance for such factors.~~

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[1,2,3]. DNase-seq allows a nucleotide-level identification of transcription factor binding sites (TFBSs). This can be performed by the computational search of footprint-like regions with low number of DNase I cuts surrounded by regions with high number of cuts[2-3]. A number of computational footprinting methods have been proposed in the past years[4-13]. Among other applications, these methods allow the delineation of the human regulatory lexicon with millions of TFBSs over distinct cell types[4], the detection of uncharacterized transcription factor (TF) motifs indicating putative regulatory elements[4] and the study of conservation of regulatory regions across different species[14].

NGS-based data are significantly affected by biases, which are inherent to the experimental protocols used[15,16,17]. 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. He et al.[15] showed that intrinsic DNase I cleavage bias around TFBSs strongly affects the performance of a computational footprinting method (footprint score; FS) ~~on~~ in a TF-specific manner. They also indicated several TFs, such as nuclear receptors and *de novo* motifs found via computational footprinting[4], where the DNase-seq profile resembles their cleavage bias estimate. Furthermore, they indicated 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, i.e. low number of DNase-seq reads surrounding the footprint. Moreover, they also noticed that nuclear receptors have DNase-seq profiles resembling cleavage bias estimates. While both studies[7,15] show the challenges imposed by cleavage bias and residence time, there ~~has~~ have been a few attempts[7,12,15] to address these computationally.

There is no well defined gold standard for the evaluation of footprinting methods. All works so far have has used ChIP-seq of TFs in conjunction with motif-based predictions as ground truth. In short, motif-predicted binding sites (MPBSs) supported by ChIP-seq peaks are positive examples (true TFBSs), while MPBSs without ChIP-seq support are negative examples (false TFBSs)[10]. This evaluation requires TF ChIP-seq experiments to be carried out on the very same cells as the DNase-seq experiment and has a few caveats. First, TF ChIP-seq peaks are also observed in indirect binding events[4,7,12,18]. Second, they have a lower spatial resolution than DNase-seq. Therefore, false TFBSs might be regarded as true TFBSs by proximity to a real TFBS of a distinct TF[15,17]. Recently, Yardımcı et al.[12] indicated that footprint quality scores, as measured by the footprint likelihood ratio (FLR), were significantly higher in cells where the TF was expressed. This observation indicates that comparing changes in expression and quality of footprints in a pairs of cells could provide an alternative footprint evaluation measure. Finally, with the exception of a few studies[8,11,12,13], comparative analyses evaluating footprinting methods were based on ChIP-seq of few (<12) TFs and with the exception of Gusmao et. al.[8], a maximum of four competing methods were evaluated. Despite the importance of method evaluation[19], there is a clear lack of benchmark data, evaluation standards and studies performing a comprehensive analysis of computational footprinting methods.

We propose here a comprehensive evaluation ofWe evaluated 10 computational footprinting methods: Neph[4], Boyle[5], Wellington[6], DNase2TF[7], HINT[8], Centipede[9], Cuellar[10], PIQ[11], FLR[12] and BinDNase[13]. In a "ChIP-seq based approach Tthey are evaluated in their accuracy to recover TFBSs supported by 88 ChIP-seq TF experiments of two cell types (H1-hESC and K562) with the area under the receiver operating characteristic curves (AUC) and precision-recall curves (AUPR). This evaluation approach will be called "ChIP-seq based". We also propose the "FLR-Exp" a new methodology for evaluation of computational footprints by associatingwhich associates the the FLR[12] scores for footprints in two cell types with the fold-change expression of the TF. This evaluation methodology will be termed "FLR-Exp". This analysis is based on the comparison of footprints and expression of 143 TFs in H1-hESC, K562 and GM12878 cells on MPBSs of 143 TFs. We also evaluate approaches for ranking footprints, strategies for dealing with cleavage bias and the effect of TF residence time on footprint predictions.

RESULTS

Computational genomic footprinting methods

Computational footprinting methods can be broadly categorized in (1) segmentation[4-8] and (2) site-centric methods[9-13] (**Table 1**). Several segmentation methods use window search to scan DNase-seq genomic profiles with a footprint-like shape – short regions with low DNase-seq digestion between short regions with high DNase-seq digestion (Neph[4] , Wellington[6], DNase2TF[7]). Another family of segmentation methods are based on hidden Markov models (HMMs), in which the hidden states model distinct levels of DNase-seq cleavage activity around footprints (Boyle[5], HINT[8]). Site-centric methods analyze DNase-seq profiles around MPBSs and classify these sites as being either bound or unbound. Most site-centric methods are based on unsupervised statistical methods like mixture models (FLR[12]), Bayesian mixture models (Centipede[9]) and combination of Gaussian process and expectation propagation (PIQ[11]). An alternative site-centric approach is proposed by Cuellar[10], which uses DNase-seq profiles as prior distribution for the detection of MPBSs. BinDNase is a supervised site-centric method based on logistic regression[13]. We also evaluate three simple statistics to rank MPBSs: position weight matrix (PWM) bit-score[10], FS (ratio of the number of DNase-seq reads inside and around a MPBS)[4,15] and TC (number of DNase-seq reads around a MPBS)[15]. These simple approaches serve as baseline footprinting methods.

There are several other relevant characteristics for computational footprinting methods. A few methods allow the inclusion of additional genomic and/or experimental evidence like conservation scores[9], distance to transcription start sites[9] and histone modifications[8-10]. Only PIQ[11] supports the analysis of several DNase-seq data sets, i.e. experiments with replicates or time series. Another important feature is the correction of DNase-seq cleavage bias, which is only supported by DNase2TF[7] ,HINT[8] variants (HINT-BC and HINT-BCN) and FLR[9]. While HINT-BC, HINT-BCN and DNase2TF use bias statistics to pre-process DNase-seq profiles; FLR builds a "cleavage bias" model within their mixture model on in a TF-specific manner.

Most methods use base pair DNase-seq resolution as primary input[4-9,11-13]. One exception is Cuellar[10], which is based on smoothed DNase-seq signals of windows with 150 bps. Smoothing of base pair resolution profiles is performed by PIQ via the use of Gaussian process models[11]. BinDNase uses a greedy backward feature selection approach, which merges read counts of neighboring genomic positions[13]. Footprinting methods also provide statistics to rank footprint predictions. Wellington[6] and DNase2TF[7] use read count statistics to provide *p*-values for each footprint. Several site-centric approaches provide either probabilities (BinDNase[13], Centipede[9], PIQ[11]) or log-odds scores (FLR[12]) of footprints. Other methods use statistics such as FS (Neph[4]), PWM (Cuelar[10]) scores or TC (HINT[8]), to rank predicted footprints.

The availability, usability and scalability of software tools implementing the methods are also important features. Public software is available for all methods but Boyle[5] and Neph[4]. HINT[8], PIQ[11] and Wellington[6] provide web pages and self-contained software to run experiments with a single or few command line calls. Only these methods natively support standard experimental/genomic formats (DNase-seq alignments, genomic regions and/or PWMs) as input. Site centric methods Cuellar[10], BinDNase[13], Centipede[9] and FLR[12] require a single execution and input data per TF/cell, while all segmentation based methods require an execution per cell only. These site centric methods have computational demands 5 times (FLR and Cuellar) to 50 times (BinDNase and Centipede) higher then the slowest segmentation method (Wellington) on our analysis with 88 TFs and 2 cells (**Sup. Table 1**). The main features of the evaluated methods are summarized in Table 1 and described in details in the Extended Methods 1.3.

Association of TF expression with footprint quality as evaluation measure

Yardımcı et al. indicated that the FLR of candidate footprints are significantly higher in cells where the TF is being expressed[12]. We expand this idea by evaluating if differences in FLR score distribution of footprints overlapping with MPBSs on a pair of cell types is proportional to differences in the expression of the respective TFs (**Fig. 1A**). We observed high average correlation values for the majority of evaluated methods (0.79) and extremely high correlation values (> 0.9) for top performing methods on comparisons between pairs of cell types H1-hESC, K562 and GM12878 (**Fig. 1B**; **Sup. Fig. 1**). We also evaluated the use of the TC and FS metrics as quality scores instead of FLR. They had lower average correlation values (TC = 0.35, FS = 0.73; Sup. Fig. 2). We opt, therefore, to use the FLR as quality measure for footprints for this evaluation procedure. The correlation between FLR score difference and expression fold change, which we refer to as “FLR-Exp”, will be used to rank footprint methods. Highest values indicate best performance. The FLR-Exp evaluation methodology only requires expression data and is therefore more generally applicable than TF ChIP-seq based evaluation. However, differently from the TF ChIP-seq based evaluation, the FLR-Exp approach cannot evaluate footprint predictions of individual TFs.

Impact of experimental bias

To understand the nature of experimental bias from DNase-seq experiments, we performed a clustering analysis on bias estimates of all 61 ENCODE Tier 1 and 2 DNase-seq data sets (**Fig 2; Sup. Table 2**). The bias correction of proteinized experiments follows the 6-mer scheme[13] and only considers reads inside DNase I hypersensitivity sites (DHSs) (Extended Methods 1.2). This strategy, which we call “DNase-seq experimental bias”, captures DNase I cleavage, read fragmentation, local chromatin structure and sequence complexity bias of DHSs. We also estimated bias on deproteinized (naked chromatin) DNase-seq experiments[13]. This strategy captures “DNase-seq cleavage bias”. We observed two major clusters discriminating “DNase-seq experimental bias” estimates from distinct DNase-seq protocols (single-hit[2] *vs* double-hit[3];, Fig. 2). Note, however, that there is a moderate correlation between bias estimates of the two clusters (average r = 0.39). DNase-seq cleavage bias estimates from the three deproteinized experiments form a sub-cluster together with experiments from double-hit protocol. This indicates that DNase-seq experimental bias is protocol-specific and differs from DNase-seq cleavage bias.

Next, we extended the analysis by He et al.[13] to evaluate the influence of cleavage bias on all evaluated footprinting methods based on the AUC at 10% false positive rate (FPR). HINT was evaluated with DNase-seq signals corrected with either “DNase-seq experimental bias” (HINT bias-corrected; HINT-BC) and

"DNase-seq cleavage bias" strategies (HINT bias-corrected on naked DNase-seq; HINT-BCN). Our analysis shows that only six out of 14 evaluated methods (Wellington, Neph, Boyle, DNase2TF, Centipede and FS) present a significant negative Pearson correlation (r = -0.35, -0.32, -0.28, -0.28, -0.24 and -0.22, respectively) between their accuracy performance and amount of DNase-seq cleavage bias (Fig. 3a; adjusted *p*-value < 0.05). Equivalent results are also observed on the same TFs and cellular conditions analyzed in He et al.[13] (**Sup. Fig. 3**). Methods explicitly using 6-mer cleavage bias statistics (HINT-BC, HINT-BCN and FLR) or performing smoothing (Cuellar, BinDNase and PIQ) are not significantly influenced by cleavage bias. Moreover, the performance of HINT-BC is the least affected by cleavage bias (r = -0.06). Pairwise comparison of AUC at 10% FPR values of all three HINT variants (HINT-BC, HINT-BCN and HINT) indicates significant gain in all predictions with bias correction (adjusted *p*-value < 10^{-30}; **Sup. Fig. 4a**). There is no significant difference between HINT-BC and HINT-BCN, but we observe a higher AUC on HINT-BC on all but 7 TFs. This indicates an advantage of the "DNase-seq experimental bias" correction for the footprint prediction problem.

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. 2b and c**; **Sup. Fig. 5**). The NRF1 and EGR1 DNase-seq profiles indicate that the bias-corrected signal fits better their sequence affinity than the uncorrected signal. We observe that k-mers with high DNase-seq experimental bias have a high CG content (r > 0.8 in 11 out of 12 cell types; **Sup. Fig. 6**). However, there is no significant correlation between CG content of MPBSs, AUC values or differences of AUC from HINT-BC, HINT-BCN and HINT (*p*-value > 0.05; **Sup. Fig. 4b**).

Comparative analysis of footprinting methods

It has been previously suggested that the TC, which is one of the simplest models to predict footprints, has better[13] or similar performance than sophisticated footprinting methods[8,12,13]. This lead us to the hypothesis that TC can be used a metric for ranking footprints. We therefore evaluated the use of TC as the ranking strategy instead of each method's own ranking for BinDNase, Centipede, Cuellar, DNase2TF, FLR, PIQ and Wellington. Previous to ranking by TC, site-centric methods required the definition of a minimum probability score to define active footprints. In all cases, using TC yielded higher AUC values (10% FPR) than using their intrinsic ranking metric (**Sup. Fig. 7**). Concerning site-centric methods, the probability of 0.9 yielded highest AUCs, with exception of BinDNase (best at 0.8). These parameters will be used in the next evaluation analyses.

We next evaluated all the competing methods by measuring the AUC at 1%, 10% and 100% FPRs using the TF ChIP-seq data. AUC at lower FPRs favors methods with higher sensitivity in expense of specificity. We also estimated the AUPR, which is indicated for cases with imbalance of positive and negative examples[20], and the FLR-Exp metric. Interestingly, all TF ChIP-seq based metrics indicate a very similar ranking (r > 0.98; **Fig. 4b**). There is also a high agreement between FLR-Exp and other metrics (r > 0.88). HINT-BC has the highest FLR-Exp, AUC and AUPR values and significantly outperforms all methods with the exception of HINT-BCN (adjusted *p*-value < 0.01; Sup. Fig. 8; Sup. Tables 3-6). Ignoring HINT variants, the next top performing method is DNase2TF, which significantly outperforms all other methods with the exception of PIQ (adjusted *p*-value < 0.01). PIQ outperforms all of its lower ranked competitors but Wellington with AUC (1% FPR) and AUPR (adjusted *p*-value < 0.01). Concerning the performance of TC, we observe that the AUC values for 10% and 100% FPR are very close to other footprinting methods (**Fig. 4A, Sup. Fig. 8**). This is not the case for AUC at 1% FPR or AUPR values. With the latter statistics, all methods but Centipede and Cuellar have significant superior performance than TC (p-value < 0.01; **Sup. Tables 3-6**).

Footprint predictions and transcription factor residence time

Despite the high average prediction values of top performing footprint methods, they consistently perform worst in a similar set of TFs, i.e. HINT-BC, DNase2TF and PIQ have 89% of TFs in common in the lower quartile of AUC at 10% FPR (Sup. Dataset 1). This list includes nuclear receptors, which has low residence binding time[7] and display a lower DNase I cleavage protection pattern (**Sup. Fig. 9**). To further investigate

this, we propose a statistic inspired by the concepts presented in Sung et al.[7] to detect TFs with potential short residence time. The protection score measures the difference between the amounts of DNase I digestion in the flanking regions and within the TFBS on DNase-seq signals corrected for experimental bias. We use this statistic to analyze the predictive performance of methods on TFs with distinct residence time. For this, we used the comprehensive data set with 233 combinations of DNase-seq experiments and TFs (Extended Methods 1.4).

We observed that TFs with known short residence time on DNA, such as nuclear receptors AR[21], ER[22] and GR[23], present a negative protection score (Fig. 5a). TFs with intermediate and long residence time on DNA (C-jun[24] and CTCF[25], respectively) present a positive protection score. The amount of protection is clearly reflected in the bias-corrected DNase-seq profiles (**Fig. 5b-d**). In addition, Fig. 5a also reveals an association of the protection score and the AUC of HINT-BC. Overall, the protection score positively correlates 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 TFs with potential short residence time can be poorly detected via DNase-seq footprints.

DISCUSSION

Our comparative evaluation analysis indicates the superior performance (in decreasing order) of HINT, DNase2TF and PIQ in the prediction of active TFBSs in all evaluated scenarios. Moreover, tools implementing these methods were user friendly and had lower computational demands than other evaluated methods. Clearly, the choice of computational footprinting approaches should also be based on experimental design aspects. For example, PIQ is the only method supporting analysis of replicates and time-series. On the other hand, studies requiring footprint predictions for latter *de novo* motif analysis should use segmentation-based approaches as HINT or DNase2TF. In contrast to positive evaluations of the TC by previous works[13,14], we show that it has poor sensitivity performance as indicated by the AUC at low FPR levels. On the other hand, we showed that the TC is the best strategy to rank footprint predictions from other methods.

The refined DNase-seq protocol and DNase I 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. The correction of DNase-seq signal on an experiment-specific manner virtually removes the effects of the experimental and cleavage biases on computational footprinting. We demonstrated that such correction can be performed prior to the execution of the computational footprinting method. On the other hand, ignoring cleavage bias might lead to false predictions, as observed previously for predicted de-novo motifs (Sup. Fig. 10). Moreover, the simple protection score can indicate footprints of TFs with potential short binding time. Thus, footprint predictions of TFs with low protection score should be interpreted with caution.

The assessment of footprint methods is a demanding task, both computationally and technically. We have created a fair and reproducible benchmarking data set for evaluation of protein-DNA binding using two validation approaches: TF ChIP-seq based and FLR-Exp. Although the rationales of the ChIP-seq based and FLR-Exp evaluation procedures are, in principle, very different, we observed a high agreement between their respective ranking of methods. This is evidence that this study provides a robust map of the accuracy of state-of-the-art computational footprinting methods. Finally, this study provides all statistics, basic data and scripts to evaluate future computational footprinting methods. This is an important resource for increasing transparency and reproducibility of research on computational methods for DNase-seq data.

Methods and Supplementary Information: Supplementary information regarding methods, computational experiments and further results are attached. Software, benchmarking data, bias estimates and further graphical results are available at 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., M.A. 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. An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414), 57-74 (2012).
[2] Crawford, G.E. et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Research 16(1), 123-131 (2006).
[3] Sabo, P.J. et al. Genome-wide identification of DNase I hypersensitive sites using active chromatin sequence libraries. PNAS 101(13), 4537-4542 (2004).
[4] Neph, S. et al. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489(7414), 83-90 (2012).
[5] Boyle, A.P. et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Research 21(3), 456-464 (2011).
[6] Piper, J. et al. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. Nucleic Acids Research 41(21), e201 (2013).
[7] Sung, M.-H.H. et al. DNase footprint signatures are dictated by factor dynamics and DNA sequence. Molecular Cell 56(2), 275-285 (2014).
[8] Gusmao, E.G. et al. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. Bioinformatics 30(22), 3143-3151 (2014).
[9] Pique-Regi, R. et al. Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Research 21(3), 447-455 (2011).
[10] Cuellar-Partida, G. et al. Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics 28(1), 56-62 (2012).
[11] Sherwood, R.I. et al. Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape. Nature Biotechnology 32(2), 171-178 (2014).
[12] Yardımcı, G.G. et al. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Research 42(19), 11865-11878 (2014).
[13] Kähärä, J. & Lähdesmäki, H. BinDNase: A discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data. Bioinformatics 31(17), 2852-2859 (2015).
[14] Stergachis, A.B. et al. Conservation of trans-acting circuitry during mammalian regulatory evolution. Nature 515(7527), 365-370 (2014).
[15] He, H.H. et al. Refined DNase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. Nature Methods 11(1), 73-78 (2014).
[16] Meyer, C. & Liu, X. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. Nature Reviews Genetics, 15(11), 709-721 (2014).
[17] Park, P.J. ChIP-seq: advantages and challenges of a maturing technology. Nature Reviews Genetics 10(10), 669-680 (2009).
[18] Teytelman, L. et al. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. PNAS 110(46), 18602-18607 (2013).
[19] Editorial. The difficulty of a fair comparison. Nature Methods, 12(4), 273-273 (2015).
[20] Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning – ICML 2006, 233-240 (2006).
[21] Tewari, A.K. et al. Chromatin accessibility reveals insights into androgen receptor activation and transcriptional specificity. Genome Biology 13(10), R88 (2012).
[22] Sharp, G.D. et al. Estrogen-receptor-alpha exchange and chromatin dynamics are ligand- and domain-dependent. Journal of Cell Science 119(Pt 19), 4101-4116 (2006).
[23] McNally, J.G. et al. The glucocorticoid receptor: rapid exchange with regulatory sites in living cells. Science 287(5456), 1262-1265 (2000).
[24] Malnou, C.E. et al. Heterodimerization with different Jun proteins controls c-Fos intranuclear dynamics and distribution. The Journal of Biological Chemistry 285(9), 6552-6562 (2010).
[25] Nakahashi, H. et al. A genome-wide map of CTCF multivalency redefines the CTCF code. Cell Reports 3(5), 1678-1689 (2013).

Figure 1 | FLR-Exp evaluation Metric. (a) FLR score distribution of footprints predicted with HINT-BC overlapping with MPBSs of selected TFs. These TFs have increasing expression in K562 (red) compared with H1-hESC cells (blue). The signed KS-statistic indicates the separation of both distributions. **(b)** Scatter plot with signed KS statistic and fold change expression for 143 TFs. There is a clear association between TF expression and KS-statistic ($r = 0.97$, adjusted p-value $< 10^{-10}$). Please define the range of the boxplots

Figure 2 | Clustering of bias estimates. Ward's minimum variance clustering on pairwise Spearman correlation coefficient (R) between all ENCODE's Tier 1 and 2 DNase-seq data sets based on each 6-mer ratio between observed and expected cleavage bias within DHS. DNase-seq experimental bias is estimateds from the single-hit (red) or double hit (blue) protocols. are indicated in red, DNase-seq experimental bias estimates from the double-hit protocol are in blue and DNase-seq cleavage bias estimates (deproteinized data sets) are in yellow. We observe a high average correlation between experimental biases estimated on DNase-seq data sets originated from the same protocol: DU = 0.89; UW = 0.84. Also, lower average correlation values are observed from experimental biases estimates from different protocols: DU *vs* UW = 0.39. The group of cleavage bias estimates based on 3 deproteinized datasets have an average correlation of 0.96.

Figure 3 Effects of DNase-seq biases on methods. (a) Association between the performance of 14 footprinting methods (relative to TC performance) and their cleavage bias estimated for 88 TFs of the cell

types H1-hESC and K562. The x-axis represents the correlation between the uncorrected and bias signal (observed *vs* bias signal; OBS). The OBS is evaluated for each TF by measuring the uncorrected DNase-seq signal and the bias signal for every MPBS that overlaps a footprint from the evaluated method. Then, the Spearman correlation is evaluated between the average uncorrected and bias signals. Higher OBS values indicate higher bias. The y-axis represents the ratio between the AUC at 10% FPR for each method and the TC method; higher values indicate higher accuracy. **(b-c)** Average bias signal (top) and uncorrected/corrected DNase-seq signal (bottom) for the TFs 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 4 | Evaluation of footprinting methods. (a) FLR-Exp values (for all pairwise comparison within cell types H1-hESC, K562 and GM12878) and median AUC (at 100%, 10% and 1% FPR) and AUPR values and **(b)** average rankings for the 15 evaluated methods. HINT-BC, HINT-BCN, HINT, DNase2TF are ranked as top four methods by all evaluation metrics. All baseline methods (FS, PWM and TC) are in the bottom four positions of the ranks. Note that BinDNase could not be evaluated with the FLR-Exp, as it requires ChIP-seq data for training.

Figure 5 | 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 n~~Nuclear receptors AR, ER and GR (short residence time, red); ~~in blue experiments with C-jun~~ (intermediate residence time, blue); ~~in green experiments with~~ CTCF (long residence time, green) and ~~in gray experiments with~~ either high protection score (> 6) or low AUC values (< 0.8) grey. **(b-d)** Average bias signal (top) and uncorrected/corrected DNase-seq signal (bottom) for the TFs 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.

Table 1 | Overview of methods. ~~We list here the m~~Main characteristics of the evaluated methods. ~~Methods are characterized by their type (SC – site centric *vs* SEG – segmentation approach), algorithm, bias correction strategy, resolution/smoothing strategy, method for ranking footprints, availability and usability. Concerning availability, m~~Methods obtain a '+' for availability if they are public available ~~('–' otherwise)~~. Boyle method is not public, but authors provide footprint predictions of a few cells. The code for Neph was obtained upon request to authors. Concerning usability, methods natively supporting standard genomic files and being executed with few commands (≤3) have '+' ('–' otherwise).

Name	Type	Algorithm	Bias Correction	Resolution Smoothing	Footprint Ranking	Availability	Usability	Others
BinDNase	SC	Logistic regression	no	bp / sliding window	probability	+	-	Require TF ChIP-seq for training
Boyle	SEG	HMM	no	bp	none	-	-	
Centipede	SC	Bayesian mix. model	no	bp	probability	+	-	Integrates histone and sequence data
Cuellar	SC	Weighted motif match	no	sliding window	PWM score	+	-	
DNase2TF	SEG	Sliding window	4-mer	bp	*p*-values	+	+	
FLR	SC	Mixture model	6-mer	bp	log-odds	+	-	Bias correction for each TF
HINT-BC	SEG	HMM	6-mer	bp	TC	+	+	Integrates histones

Neph	SEG	Sliding window	no	bs	FS	-	-	
PIQ	SEG	GP/expectation propagation	no	bp / GP	probability	+	+	Support replicates, time series
Wellington	SEG	Sliding window	no	bp	*p*-value	+	+	