

Supplement materials for: Addressing DNase-seq cleavage bias and residence time on computational footprinting

Eduardo G. Gusmao^{1,2}, Martin Zenke^{1,2}, and Ivan G. Costa^{*1,2,3}

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

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

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

*ivan.costa@rwth-aachen.de

Supplementary Methods

Section	Title	Page
1.1	Data	4
1.2	Bias Correction	4
1.2.1	DNase I Hypersensitive Sites	4
1.2.2	Estimation of DNase I Cleavage Bias	5
1.2.3	DNase I Cleavage Bias Correction	6
1.3	Computational Footprinting Methods (in Chronological Order)	7
1.3.1	Neph Method	7
1.3.2	Boyle Method	7
1.3.3	Centipede	8
1.3.4	Cuellar Method	8
1.3.5	Wellington	8
1.3.6	Protein Interaction Quantification (PIQ)	9
1.3.7	Footprint Mixture (FLR)	9
1.3.8	DNase2TF	9
1.3.9	HINT, HINT-BC, HINT-BCN	10
1.3.10	Footprint Score (FS)	10
1.3.11	Tag Count (TC)	11
1.4	Evaluation	11
1.4.1	Motif-Predicted Binding Sites (MPBSs)	11
1.4.2	Method Comparison	11
1.4.3	Statistical Methods	12
1.5	Protection Score	12

Supplementary Figures

Item	Descriptive Title	Page
Supplementary Fig. 1	Correlation between the performance of methods and their OBS on He Dataset	15
Supplementary Fig. 2	Average DNase-seq signals around selected TFs with ChIP-seq evidence in H1-hESC (DU) cell type	16
Supplementary Fig. 3	Ward's minimum variance clustering on pairwise Spearman correlation coefficient (R) between different DNase-seq data sets	17
Supplementary Fig. 4	Association between k-mer CG content and cleavage bias	18
Supplementary Fig. 5	AUC distribution for 14 footprinting methods regarding all validation sets (ordered by median AUC)	19
Supplementary Fig. 6	Average DNase-seq signals around nuclear receptor TFs with ChIP-seq evidence in LNCaP(DU), m3134(UW) and MCF-7(DU) cell types	20
Supplementary Fig. 7	Average bias and DNase-seq signals around binding sites of <i>de novo</i> motifs 0458 and 0500 on cell type H7-hESC	21

Supplementary Tables

Item	Descriptive Title	Page
Supplementary Table 1	Summary of DNase-seq data	22
Supplementary Table 2	Summary of computational footprinting methods	23
Supplementary Table 3	Friedman-Nemenyi hypothesis test results on AUC for all evaluated methods	24

1 Methods

1.1 Data

DNase-seq aligned reads were obtained from ENCODE (ENCODE Project Consortium, 2012). We obtained data regarding cell types H1-hESC, HeLa-S3, HepG2, Huvec, K562, LNCaP and MCF-7 from Crawford’s Lab (labeled with the initials of their institution “DU”) and concerning cell types H7-hESC, HepG2, Huvec, K562 and m3134 from Stamatoyannopoulos’ lab (labeled with the initials of their institution “UW”). We also used deproteinized DNase-seq experiments from cell types MCF-7 and K562 (Crawford lab) (Yardımcı *et al.*, 2014) and IMR90 (Stamatoyannopoulos lab) (Lazarovici *et al.*, 2013). DNase-seq experiments labeled with “DU” follow the single-hit protocol, while the experiments labeled with “UW” follow the double-hit protocol. See Supplementary Table 1 for data description.

Transcription factor (TF) ChIP-seq enriched regions (peaks and summits) were obtained in ENCODE Analysis Working Group (AWG) track with exception of the following experiments, in which the enriched regions were obtained using bowtie-2 (Langmead and Salzberg, 2012) and MACS (Zhang *et al.*, 2008). AR (R1881 treatment) ChIP-seq raw sequences for LNCaP cell type was obtained in Gene Expression Omnibus (GEO) with accession number GSM353644 (Yu *et al.*, 2010). ER (40 and 160 minutes after estradiol treatment) ChIP-seq raw sequences for MCF-7 cell type was obtained in GEO with accession number GSE54855 (Guertin *et al.*, 2014). GR (dexamethasone treatment) ChIP-seq raw sequences for m3134 cell type was obtained in SRA under study number SRP004871 (John *et al.*, 2011). All organism-specific data (DNase-seq and ChIP-seq) are based on the human genome build 37 (hg19), except the DNase-seq for m3134 and ChIP-seq for GR, which were based on mouse genome build 37 (mm9). Chromosome Y was removed from all analyses.

TF motifs (position frequency matrices; PFMs) were obtained from the Jaspar (Mathelier *et al.*, 2014), Uniprobe (Robasky and Bulyk, 2011) and Transfac (Matys *et al.*, 2006) repositories. Non-organism-specific data (PFMs) were obtained for the subphylum Vertebrata. *de novo* PFMs 0458 and 0500 were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/footprints/jan2011/de.novo.pwm (Neph *et al.*, 2012). The accession codes for all TF ChIP-seq experiments and PFM IDs are available in the Supplementary Dataset 1.

1.2 Bias Correction

1.2.1 DNase I Hypersensitive Sites

A first task is the identification of DNase I Hypersensitivity sites (DHSs). A nucleotide-resolution genome-wide signal was created for each DNase-seq data set by counting reads mapped to the genome. Here, we considered only the 5’ position of the aligned reads (position at which DNase I cleaved the DNA). The genomic signal was created by counting the

number of reads that overlapped at each genomic position.

More formally, we define a raw genomic signal as a vector

$$\mathbf{x} = \langle x_1, \dots, x_N \rangle,$$

where N equals the number of bases in the genome and each $x_i \in \mathbb{N}^0$ is the number of DNase-seq reads mapped to position i . We also generate strand specific counts X^s , where $s \in \{+, -\}$ describes the strand the read was mapped to.

DNase I hypersensitivity sites are estimated based on the DNase I raw signal. First, the F-seq software (Boyle *et al.*, 2008) was used to create smoothed DNase-seq signals using Parzen density estimates. Then, the smoothed signal \mathbf{x}^{fseq} was fit to a gamma distribution,

$$\mathbf{x}^{\text{fseq}} \sim \Gamma(\kappa, \theta),$$

by evaluating κ and θ based on mean and standard deviation estimates. Finally, the enriched regions (DHSs) were found by establishing a cutoff based on a p -value of 0.01 (Boyle *et al.*, 2008). We refer to DHSs as a set of genomic intervals

$$H = \{h_1, \dots, h_L\},$$

where $h_i = [m, n]$ for $m < n \in \mathbb{N}$ and L is the total number of DHSs ¹.

1.2.2 Estimation of DNase I Cleavage Bias

We use two approaches to estimate of the intrinsic DNase I cleavage bias: (1) aligned reads inside DHSs from DNase-seq experiments (termed “DHS approach”) and (2) all aligned reads for deproteinized (naked) DNA experiments (termed “naked DNA approach”). The observed cleavage score for a k -mer w corresponds to the number of DNase I cleavage sites centered on w . The background cleavage score is defined by the total number of times w occurs. Then, the bias estimation is computed as the ratio between the observed and background cleavage scores. **Mathematical formalizations of the bias estimation will be made based on the DHS approach.**

We define G^s as the reference genome sequence with length N for strand $s \in \{+, -\}$. $G^s[i..j]$ indicates the sequence from positions i to j (including both within the interval). For each k -mer w with length k the observed cleavage score o_w can be calculated as

$$o_w^s = 1 + \sum_{i=1}^L \sum_{j \in h_i} x_j^s \mathbf{1} \left(G^s \left[j - \frac{k}{2} .. j + \frac{k}{2} \right] = w \right), \quad (1)$$

where $\mathbf{1}(\cdot)$ is an indicator function.

Similarly, the background cleavage score r_w can be evaluated as

$$r_w^s = 1 + \sum_{i=1}^L \sum_{j \in h_i} \mathbf{1} \left(G^s \left[j - \frac{k}{2} .. j + \frac{k}{2} \right] = w \right). \quad (2)$$

¹We ignore for simplicity of notation the fact that intervals are defined on distinct chromosomes or contigs

Finally, the cleavage bias b_i^s for a genomic position $k + 1 \leq i \leq N - k + 1$, given that $w = G^s[i - \frac{k}{2}..i + \frac{k}{2}]$, can be calculated as

$$b_i^s = o_w^s \cdot R / r_w^s \cdot O^s, \quad (3)$$

where O^s indicates the total number of reads aligned to strand s in DHSs

$$O^s = \sum_{i=1}^L \sum_{j \in h_i} x_j^s, \quad (4)$$

and R indicates the total number of k -mers in DHS positions

$$R = \sum_{i=1}^L \sum_{j \in h_i} 1. \quad (5)$$

The bias score b_i^s represents how many times the k -mer sequence $G^s[i - \frac{k}{2}..i + \frac{k}{2} + 1]$ was cleaved by the DNase I enzyme in comparison to its total occurrence in: (1) DHSs (DHS approach); (2) the entire genome (naked DNA approach). As observed by He *et al.* (2014) a 6-mer bias model captures more information than $k < 6$ models and the information added with $k > 6$ models are not significant. Therefore, in this study, all analyses were performed using a 6-mer bias model.

1.2.3 DNase I Cleavage Bias Correction

A “smoothed corrected signal” was calculated using smoothed versions of both raw DNase-seq (\hat{x}_i^s) and the bias score signal (\hat{b}_i^s) (He *et al.*, 2014). These smoothed signals were based on a 50 bp window and can be written as

$$\hat{x}_i^s = \sum_{j=i-25}^{i+24} x_j^s \quad \hat{b}_i^s = \frac{b_i^s}{\sum_{j=i-25}^{i+24} b_j^s}. \quad (6)$$

With these results we are able to define the smoothed corrected signal as

$$c_i^s = \hat{x}_i^s \hat{b}_i^s. \quad (7)$$

Finally, the bias-corrected DNase-seq genomic signal (y) can be obtained by applying

$$y_i^s = \log(x_i^s + 1) - \log(c_i^s + 1). \quad (8)$$

The corrected DNase-seq signal generated by Eq. 8 may include negative values. Since some posterior statistical analyses required a signal consisting only of positive values, we have shifted the entire signal by adding the global minimum value.

1.3 Computational Footprinting Methods (in Chronological Order)

In this section we present an overview of the computational footprinting methods used in this study. Also, we provide a detailed description of the parameterization of each method. The Supplementary Table 2 shows a summary of all methods evaluated in this study according to main features from footprinting methods.

1.3.1 Neph Method

Neph *et al.* (2012) used a simplified version of the segmentation-based method originally proposed in Hesselberth *et al.* (2009). Their method consists on applying a sliding window to find genomic regions (6–40 bp) with low DNase I cleavage activity between regions (3–10 bp) with intense DNase I digestion. A footprint occupancy score (FOS) is evaluated and used to determine the most significant predictions.

We obtained the footprint predictions for cell type K562 (DU) in ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/footprints/jan2011/all.footprints.gz. As predictions were not available for H1-hESC (DU), we obtained the scripts and parameterization through personal communication with S. Neph. Briefly, we used the DNase I raw signal as input with the parameters from the original publication: flanking component length varied between 3–10 bp and central footprint region length varied between 6–40 bp. Afterwards, the footprints were filtered by a $FOS \leq 0.95$, which generally agrees with a false discovery rate (FDR) of 1%. Finally, we consider only predictions that occurred within DNase-seq hotspots, evaluated using the method first described in Sabo *et al.* (2004). The DNase-seq hotspot for K562 are provided in ENCODE (EwgEncodeEH000480; GSM646567). We will refer to this framework as “Neph”.

1.3.2 Boyle Method

Boyle *et al.* (2011) designed a segmentation-based approach, which is based on using hidden Markov models (HMMs) to predict footprints in specific DNase I cleavage patterns. Briefly, the HMM uses a normalized DNase-seq cleavage signal to find regions with depleted DNase I digestion (footprints) between two peaks of intense DNase I cleavage. Such pattern reflects the inability of the DNase I nuclease to cleave sites where there are proteins bound. As the DNase-seq profiles required a nucleotide-resolution signal, which is usually noisy, the authors used a Savitzky-Golay smoothing filter to reduce noise and to estimate the slope of the DNase-seq signal (Madden, 1978). Their HMM had five states, with specific states to identify the decrease/increase of DHS signals around the peak-dip-peak region. Since no source code or software is provided, we used footprint predictions from Boyle *et al.* (2011) available at <http://fureylab.web.unc.edu/datasets/footprints/>. We will refer to this method as “Boyle”.

1.3.3 Centipede

Centipede is a site-centric approach that consists on gathering experimental and genomic information around MPBSs and using an unsupervised Bayesian approach to label each retrieved site as ‘bound’ or ‘unbound’ (Pique-Regi *et al.*, 2011). The experimental and genomic data used include DNase-seq, position weight matrix (PWM) bit-score, sequence conservation and distance to the nearest transcription start site (TSS). The experimental data input was generated by fetching the raw DNase-seq signal surrounding a 200 bp window centered on each MPBS. Additionally, to create the genomic data input, we obtained PhastCons conservation score (placental mammals on the 46-way multiple alignment) (Siepel *et al.*, 2005) and Ensembl gene annotation from ENCODE (Hubbard *et al.*, 2002) to create the prior probabilities in addition to the PWM bit-score.

Centipede software was obtained at <http://centipede.uchicago.edu/> and executed to generate posterior probabilities of regions being bound by TFs. We have previously observed that Centipede is sensitive to certain parameters. Therefore, Centipede parameterization was defined with an extensive computational evaluation described in Gusmao *et al.* (2014).

1.3.4 Cuellar Method

Cuellar-Partida *et al.* (2012) proposed a site-centric method to include DNase-seq data as priors for the detection of active transcription factor binding sites (TFBSs). It is based on a probabilistic classification approach to compute better log-posterior odds score than the ones observed by purely sequence-based approaches. We applied this method as described in Cuellar-Partida *et al.* (2012). We created a smoothed DNase-seq input signal by evaluating the number of DNase-seq cleavage based on a 150 bp window with 20 bp steps. We obtained their scripts at <http://research.imb.uq.edu.au/t.bailey/SD/Cuellar2011/> and created priors using the smoothed version of the DNase-seq signal. As suggested by the authors, the priors were submitted to the program FIMO (Grant *et al.*, 2011) to obtain the predictions. We will refer to this method as “Cuellar”.

1.3.5 Wellington

Wellington is a segmentation approach based on a Binomial test. For a given candidate footprint, it tests the hypothesis that there are more reads in the flanking regions than within the footprint. Following an observation that DNase-seq cuts of the double-hit protocol are strand-specific, Wellington only considers reads mapped to the upstream flanking region of the footprints. Wellington automatically detects the size of footprints (within a user-defined interval) and sets flanking regions at a user-defined length. We have obtained Wellington’s source code in <http://jpiper.github.com/pyDNase> and executed it with default parameters. Briefly, we used a footprint p -value cutoff of -30 , footprint sizes varying between 6 and 40 with 1 bp steps and shoulder size (flanking regions) of 35 bp.

1.3.6 Protein Interaction Quantification (PIQ)

The Protein Interaction Quantification (PIQ) is a site-centric method, which uses Gaussian process to model and smooth the footprint profiles around candidate MPBSs (± 100 bp) (Sherwood *et al.*, 2014). Active footprints are estimated with an expectation propagation algorithm. Finally, PIQ indicates the set of motifs which footprint signals are distinguishable from noise to reduce the set of candidate transcription factors. We obtained PIQ implementation in <http://piq.csail.mit.edu> and executed it with default parameters, which can be found in the script `common.r`. Briefly, putative binding sites were generated with the script `pwmmatch.exact.r`. The DNase-seq signal was created using the script `bam2rdata.r`. And the footprints were detected with the script `pertf.r`.

1.3.7 Footprint Mixture (FLR)

Yardımcı *et al.* (2014) proposed a site-centric method based on a mixture of multinomial models to detect active/inactive MPBSs. The method uses an expectation maximization algorithm to find a mixture of two multinomial distributions, representing active (footprints) and inactive (background) MPBSs. The background model is initialized with either bias cleavage frequencies or estimated *de novo*. After successful estimation, MPBSs are scored with the log odds ratio for the footprint vs. background model. The model takes DNase-seq cuts within a small window around the candidate profiles (± 25 bp) as input. DNase-seq cleavage bias is estimated for 6-mers based on the DNA sequences extracted within the same regions in which the cuts were retrieved. Method implementation was obtained in https://ohlerlab.mdc-berlin.de/software/FootprintMixture_109/. We executed the method using cleavage bias frequencies for initialization of the background models. The width of the window surrounding the binding site (*PadLen*) was set to the default value of 25 bp. Also, we use the expectation maximization to re-estimate background during training (argument *Fixed* set to `FALSE`). We will refer to this method as “FLR”.

1.3.8 DNase2TF

DNase2TF is a segmentation-based approach based on a binomial z-score, which evaluates the depletion of DNase-seq reads around the candidate footprints Sung *et al.* (2014). At a second step, DNase2TF interactively merges close candidate footprints whenever they improve depletion scores. DNase2TF corrects for DNase cleavage bias using cleavage statistics for 2 or 4-mers. We obtained source code from <http://sourceforge.net/projects/dnase2tfr/> and executed DNase2TF with a 4-mer cleavage bias correction. Other parameters were set to their default values: $minw = 6$, $maxw = 30$, $z_threshold = -2$ and $FDR = 10^{-3}$.

1.3.9 HINT, HINT-BC, HINT-BCN

Recently, Gusmao *et al.* (2014) have proposed the segmentation method HINT (HMM-based identification of transcription factor footprints) as an extension of Boyle method (Boyle *et al.*, 2011). HINT is based on eight-state multivariate HMMs and combines DNase-seq and histone modification ChIP-seq profiles at the nucleotide level for the identification of footprints. The pipeline of HINT method starts by normalizing the DNase I cleavage signal using within- and between-dataset normalizations. Then, the slope of the normalized signals is evaluated to identify the DNase-seq signal increase and decrease. Afterwards, an HMM is trained on a supervised manner (maximum likelihood) based on manually annotated genomic regions. To aid such manual annotation the normalized and slope signals are used in combination with MPBSs for all available PFMs in the repositories Jaspar and Uniprobe. Finally, the Viterbi algorithm is performed on the trained HMMs inside regions consisting of DHSs extended by 5,000 bp upstream and downstream. All parameters were set as described in Gusmao *et al.* (2014).

We have performed two modifications to the method described in Gusmao *et al.* (2014). First, to perform a standardized comparison, we modified HINT to allow only DNase-seq data. The modified HMM model contains five states. The three histone-level states were removed and new transitions were created from the **BACKGROUND** state to the DNase **UP** state and from the DNase **DOWN** state to the **BACKGROUND** state. The second modification concerns the use of bias-corrected DNase-seq signal prior to normalization steps. We will call the method HINT bias-corrected (HINT-BC), for correction based on the DHS approach, and HINT bias-corrected naked DNA (HINT-BCN), for bias correction based on the naked DNA approach. These modifications required retraining of the HMM models. For this, we used the same manual annotation described in Gusmao *et al.* (2014). The novel methods and trained models are available as a command-line tool at www.costalab.org/hint-bc.

1.3.10 Footprint Score (FS)

He *et al.* (2014) used a site-centric MPBS ranking scheme termed “footprint score (FS)”, which is based on a scoring metric from the footprinting methodology proposed in Neph *et al.* (2012). The FS statistic is defined as

$$\text{FS}_{\text{MPBS}_i} = - \left(\frac{n_{C,i} + 1}{n_{R,i} + 1} + \frac{n_{C,i} + 1}{n_{L,i} + 1} \right), \quad (9)$$

where $\text{MPBS}_i = [m_i, n_i]$ is the i -th MPBS which extends from genomic positions m_i to n_i and $\overline{\text{MPBS}_i} = (m + n)/2$. The FS uses the DNase-seq signal in the center ($n_{C,i}$) of the MPBS and its upstream ($n_{L,i}$) and downstream ($n_{R,i}$) flanking regions. These variables can be defined as

$$n_{C,i} = \sum_{j=m_i}^{n_i} x_j, \quad n_{R,i} = \sum_{j=n_i}^{2n_i-m_i} x_j, \quad n_{L,i} = \sum_{j=2m_i-n_i}^{m_i} x_j. \quad (10)$$

1.3.11 Tag Count (TC)

The site-centric method which we refer to as “tag count (TC)”, corresponds to the number of DNase I cleavage hits in a 200 bp window around predicted TFBS as defined in He *et al.* (2014). This can be written as

$$TC_{MPBS_i} = \sum_{j=\overline{MPBS_i}-100}^{\overline{MPBS_i}+99} x_j. \quad (11)$$

1.4 Evaluation

1.4.1 Motif-Predicted Binding Sites (MPBSs)

Method evaluation was performed with a site-centric binding site statistics. For this, we generated position weight matrices (PWMs) from PFMs by evaluating the information content of each position and performing background nucleotide frequency correction (Stormo, 2000). This was performed using Biopython (Cock *et al.*, 2009). Then, we created motif-predicted binding sites (MPBSs) by matching all PWMs against the human(hg19)/mouse(mm9) genome using the fast performance motif matching tool MOODS (Korhonen *et al.*, 2009). This procedure produces “PWM bit-scores” for every match. We determined a bit-score cutoff threshold by applying the dynamic programming approach described in Wilczynski *et al.* (2009) with a false positive rate (FPR) of 10^{-4} . All site-centric scores were based on the set of MPBSs after the application of the cutoff threshold. Also, the PWM bit-score was used as a control metric and will be referenced as “PWM”.

1.4.2 Method Comparison

All methods were evaluated using a site-centric approach (Cuellar-Partida *et al.*, 2012), which combines MPBSs with ChIP-seq data for every TF. In this scheme, MPBSs with ChIP-seq evidence (located within 100 bp from the ChIP-seq peak summit) are considered “true” TFBSs; while MPBSs without ChIP-seq evidence are considered “false” TFBSs. Every TF prediction that overlaps a true TFBS is considered a correct prediction (true positive – TP) and every prediction that overlaps with a false TFBS is considered an incorrect prediction (false positive – FP). Therefore, true negatives (TN) and false negatives (FN) are, respectively, false and true TFBSs without overlapping predictions.

To assess the accuracy of digital genomic footprinting methods we created receiver operating characteristic (ROC) curves. Briefly, these curves describe the sensitivity increase as we decrease the specificity of the method. Furthermore, the area under the ROC curve (AUC) metric was evaluated at the 10% false positive rate (FPR). **Segmentation-based approaches (Boyle, DNase2TF, HINT, Neph and Wellington)** provide footprint predictions that do not necessarily encompass all MPBSs. To create full ROC curves for these methods, we first ranked all predicted

sites by their DNase I cleavage tag count followed all non-predicted sites ranked by their tag count. In order to present a fair comparison, this approach was also applied to all site-centric methods (Centipede, Cuellar, FLR and PIQ). For that, we considered a probability threshold of 0.9 for detection of footprint on all site-centric methods.

Our comparative experiments comprise the following three evaluation scenarios. All evaluation statistics and method performances are available at the Supplementary Dataset 1.

He Dataset: To replicate the analysis performed in He *et al.* (2014), we analyzed DNase-seq from cell types K562(UW), LNCaP(DU) and m3134(UW) on 36 TFs and we evaluated the methods PWM, FS, TC, HINT, HINT-BC and HINT-BCN.

Benchmarking Dataset: For comparative analysis of several competing methods, we selected the two cell types with highest number of transcription factor ChIP-seq data sets evaluated in our study: K562(DU) with 59 factors and Hihesc(DU) with 29 factors. We can therefore make use of predictions provided by Gusmao *et al.* (2014) and Boyle *et al.* (2011), which includes evaluation of Boyle, Cuellar, Centipede, HINT and Neph methods. For this data set, we have estimated novel footprints for FS, TC, HINT-BC, HINT-BNC, DNase2TF, PIQ, Wellington and FLR methods, which were not previously evaluated.

Comprehensive dataset: Lastly, we have compiled a comprehensive data set containing 233 combinations of cells and transcription factors with matching cellular background. This data set was built from a catalog of 144 TF ChIP-seq and 13 DNase-seq data sets. This data is used to evaluate the effects of bias correction and transcription factor binding time. In this scenario we evaluated the methods PWM, FS, TC, HINT, HINT-BC and HINT-BCN.

1.4.3 Statistical Methods

The non-parametric Friedman-Nemenyi hypothesis test (Demšar, 2006) was used to compare the AUC of the methods regarding all data set combinations (TFs *vs.* cell types). Such test provides a rank of the methods as well as the statistical significance of whether a particular method was outperformed. Statistical evaluation of correlation values were based on two-sided t-test provided by R function `corr.test`. All reported *p*-values have been corrected with the Benjamini and Hochberg method (Benjamini and Hochberg, 1995).

1.5 Protection Score

We propose a measure to detect TF-specific footprint protection for a given DNase-seq experiment and MPBSs of a given motif/TF. As previously indicated in Sung *et al.* (2014), fewer DNase-seq cuts (protection) surrounding the binding site characterizes transcription factors with shorter binding times. More formally, the protection score for a set of **MPBS** is defined

as:

$$\text{PROT}_{\text{MPBS}} = \sum_{i=1}^N \frac{(n_{R,i} - n_{C,i}) + (n_{L,i} - n_{C,i})}{2N}, \quad (12)$$

where $\text{MBPS} = \{\text{MPBS}_1, \dots, \text{MPBS}_N\}$ is set of binding sites for a given motif, $\text{MPBS}_i = [m_i, n_i]$ is the genomic location of the i th binding site and $n_{C,i}$, $n_{L,i}$, $n_{R,i}$ are the number of DNase reads in the binding site, upstream and downstream flanking positions, respectively (see Eq. 10 for details).

In short, the protection score indicates the average difference of DNase-seq counts in the flanking region and the DNase-seq counts within the MPBS. Positive values will indicate protection in the flanking regions, while values close to zero or negative indicates no protection. The protection score is a similar statistic as the Footprint Score (FS) (Sec. 1.3.10). The main difference is that the FS score measures the ratio between reads in flanking vs. binding sites, while the protection score measures the difference. Finally, since we are interested in using the protection score as a measure of quality for a given transcription factor and set of footprint predictions, we only evaluate MPBSs overlapping with footprints for a given cell type. The DNase-seq count values are previously corrected for cleavage bias and coverage differences.

References

- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289–300.
- Boyle, A. P., *et al.* (2008). F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**(21), 2537–2538.
- 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.
- Cock, P. J. A., *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.
- Cuellar-Partida, G., *et al.* (2012). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, **28**(1), 56–62.
- Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, **7**, 1–30.
- ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.
- Grant, C. E., Bailey, T. L., and Noble, W. S. (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**(7), 1017–1018.
- Guertin, M., *et al.* (2014). Transient estrogen receptor binding and p300 redistribution support a squelching mechanism for estradiol-repressed genes. *Mol Endocrinol*, **28**(9), 1522–1533.
- 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.
- 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.
- Hesselberth, J. R., *et al.* (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, **6**(4), 283–289.

- Hubbard, T., *et al.* (2002). The ensembl genome database project. *Nucleic acids research*, **30**(1), 38–41.
- John, S., *et al.* (2011). Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat Genet*, **43**(3), 264–268.
- Korhonen, J., *et al.* (2009). MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics*, **25**(23), 3181–3182.
- Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nat Meth*, **9**(4), 357–359.
- Lazarovici, A., *et al.* (2013). Probing DNA shape and methylation state on a genomic scale with DNase i. *Proceedings of the National Academy of Sciences*, **110**(16), 6376–6381.
- Madden, H. H. (1978). Comments on the Savitzky-Golay convolution method for least-squares fit smoothing and differentiation of digital data. *Anal.Chem.*, **50**, 1383–1386.
- Mathelier, A., *et al.* (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, **42**(D1), D142–D147.
- Matys, V., *et al.* (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, **34**(Database issue), D108–D110.
- Neph, S., *et al.* (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**(7414), 83–90.
- 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.
- Robasky, K. and Bulyk, M. L. (2011). UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic acids research*, **39**(Database issue).
- Sabo, P. J., *et al.* (2004). Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(48), 16837–16842.
- 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), 171–8.
- Siepel, A., *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, **15**(8), 1034–1050.
- Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, **16**(1), 16–23.
- Sung, M.-H. H., *et al.* (2014). DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Molecular cell*, **56**(2), 275–285.
- Wilczynski, B., *et al.* (2009). Finding evolutionarily conserved cis-regulatory modules with a universal set of motifs. *BMC bioinformatics*, **10**(1), 82+.
- Yardımcı, G. G., *et al.* (2014). Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic acids research*, **42**(19), 11865–78.
- Yu, J., *et al.* (2010). An integrated network of androgen receptor, polycomb, and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell*, **17**(5), 443–454.
- Zhang, Y., *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, **9**(9), R137+.

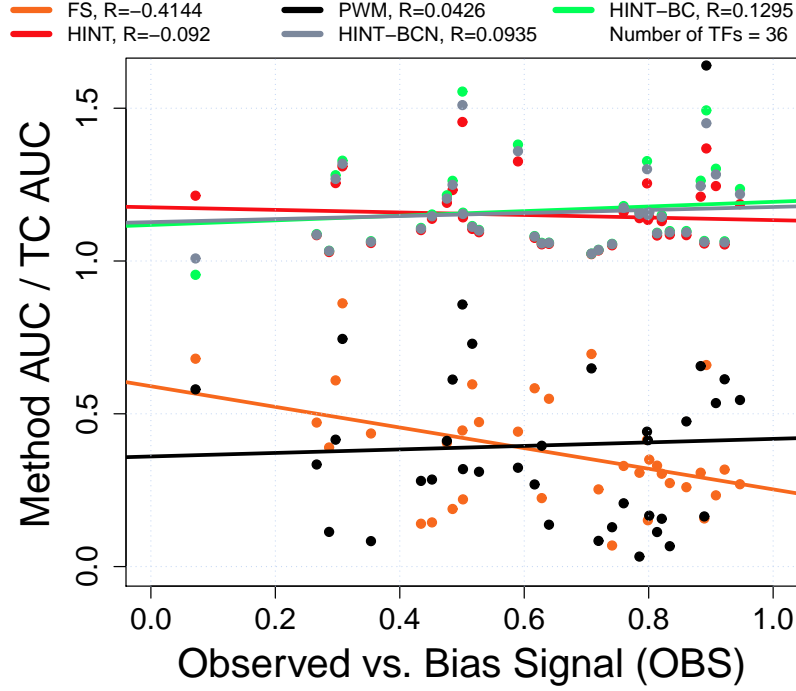
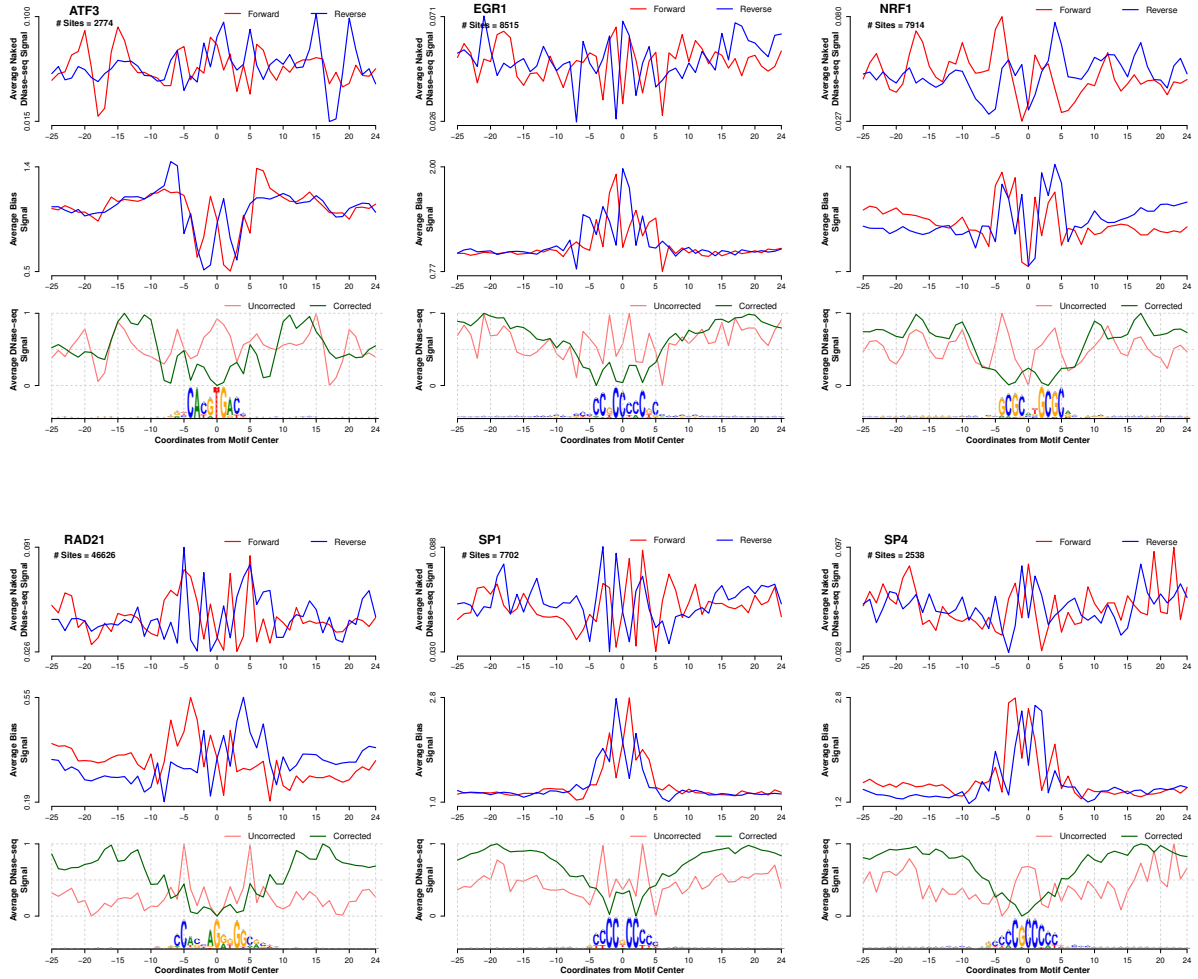


Figure 1: Correlation between the performance of methods and their OBS on He Dataset. The x -axis represents the observed sequence bias. The y -axis represents the ratio between the AUC at 10% FPR for a particular method and the TC method. In accordance with He *et al.* (2014), we observe that FS method has a high negative correlation ($R = -0.4144$; adjusted p -value < 0.001) with the cleavage bias score, while no significant correlation is found for all other evaluated methods HINT, HINT-BCN, HINT-BC and PWM. It is important to notice that the correlation value for FS method differs from He *et al.* (2014). This stems from a different strategy to find the DNase hypersensitivity regions and MPBSs used in the evaluation dataset. Nevertheless, we were able to observe a strong bias for the FS method as in He *et al.* (2014).



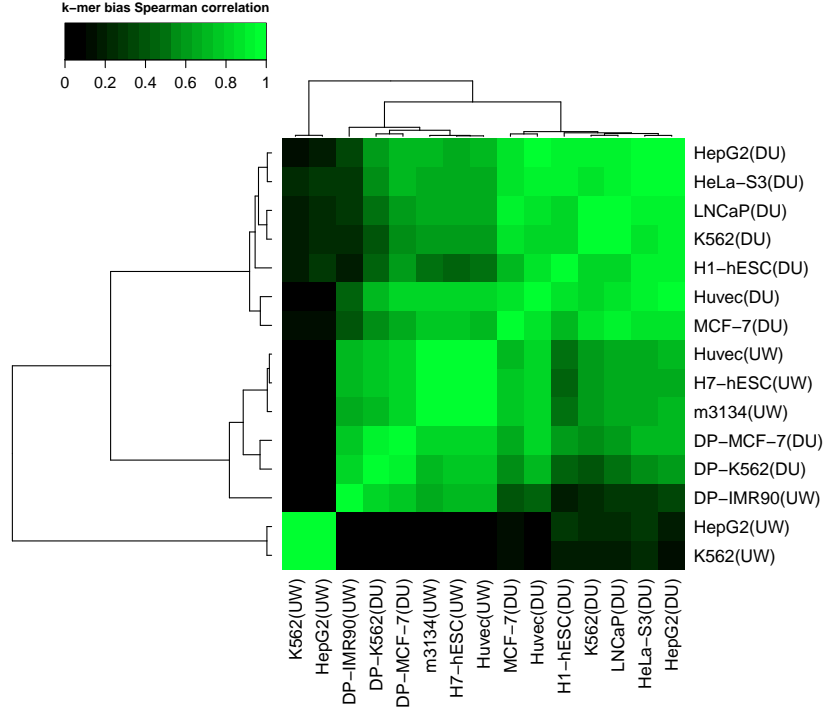


Figure 3: Ward’s minimum variance clustering on pairwise Spearman correlation coefficient (R) between different DNase-seq data sets calculated based on each 6-mer w ratio between observed and expected cleavage bias ($o_w^s \cdot R/r_w^s \cdot O^s$). Crawford’s lab data sets are represented by “DU” (Duke University; single-hit protocol) and Stamatoyannopoulos’ lab data sets are represented by “UW” (University of Washington; double-hit protocol). We observe 3 major clusters: group 1 contains all DU data sets, group 2 contains m3134, Huvec from UW and all deproteinized DNase-seq experiments (IMR90(UW); K562 and MCF-7(DU)) and group 3 contains K562 and HepG2 from UW. Correlation between experiments clustered into the same group are significant (adjusted p -value < 0.001). These results confirm that DHS-estimated cleavage bias experiments from the same lab/protocol have similar bias. The only exceptions are K562(UW) and HepG2(UW), which presented low correlation values with any other UW and DU experiments. Deproteinized DNA experiments from distinct protocols clustered together. The high correlations between deproteinized DNA experiments ($R = 0.94$ for K562 and MCF-7; $R = 0.81$ for K562 and IMR90) are in agreement with previous reports [Yardımcı *et al.* \(2014\)](#). In particular, deproteinized DNA experiments have a moderate correlation with experiments using the same protocols: K562 has an average $R = 0.42$ with other DU experiments, MCF-7 has average $R = 0.60$ with other DU experiments; while IMR90 has average $R = 0.67$ with m3134(UW) and Huvec(UW). While deproteinized DNA could be used for estimation bias correction for experiments with same protocol, we observe a lack of correlation between deproteinized IMR90(UW) and K562/HepG2(UW) experiments. One possible reason is that small variations of the same protocol might introduce further cleavage bias.

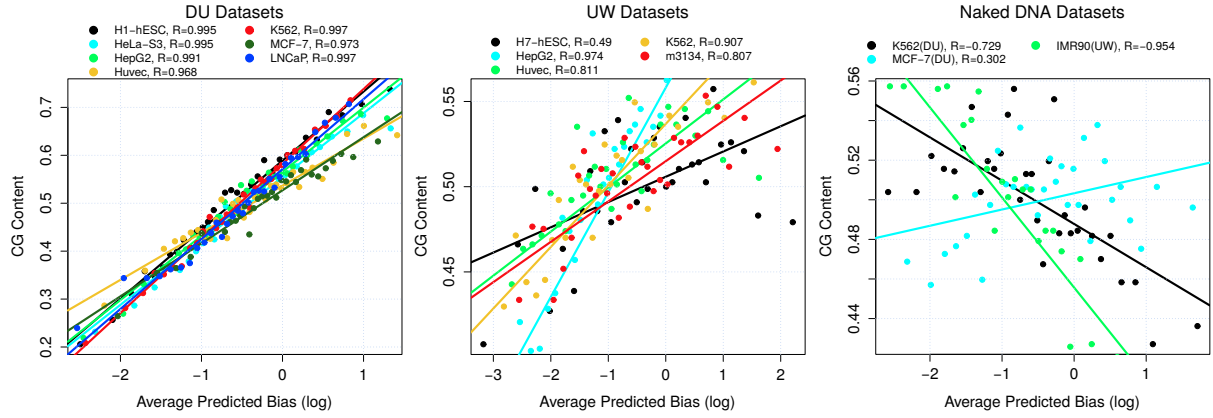


Figure 4: Association between k-mer CG content and cleavage bias. We sorted k-mers by their cleavage bias estimate and grouped similar ranked k-mers in 32 groups. We show scatter plots with CG content vs. average cleavage bias for k-mer groups on DHS-based k-mers estimated from single-hit (DU), double-hit (WU) and naked DNA experiments. There is a strong positive correlation between cleavage bias and CG content for all DHS-based estimates from both single-hit and double-hit protocols (adjusted p-value ≤ 0.01). Interestingly, we observe a negative correlation for two deproteinized DNA experiments: K562(DU) and IMR90(UW) (adjusted p-value ≤ 0.00001). These results reinforce the distinctions between naked DNA and DHS-based cleavage bias estimates. Moreover, it explains the positive impact of cleavage bias correction on TFs with CG-rich motifs, as the ones described in Supplementary Figure 2.

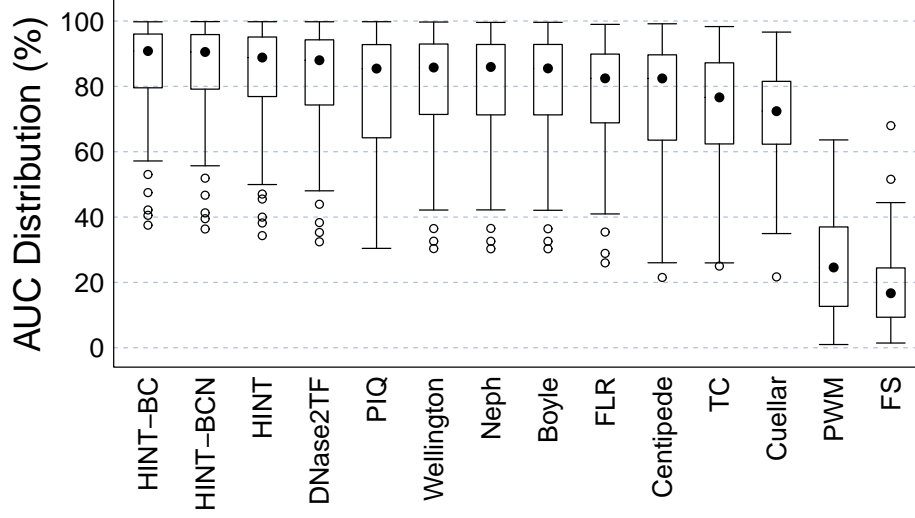


Figure 5: AUC distribution for 14 footprinting methods regarding all validation sets (ordered by median AUC). We used the Friedman-Nemenyi ranking and hypothesis test for statistical evaluation (Demšar, 2006) (Supplementary Tables 3). All segmentation-based approaches (HINT-BC, HINT-BCN, HINT, Boyle, DNase2TF and Wellington) and the site centric method PIQ significantly outperformed TC (adjusted p -value < 0.01). Moreover, HINT-BC and HINT-BCN outperformed all other competing methods (adjusted p -value < 0.01). As reported in He *et al.* (2014), the AUC of the TC approach was significantly higher than FS method (adjusted p -value < 0.01).

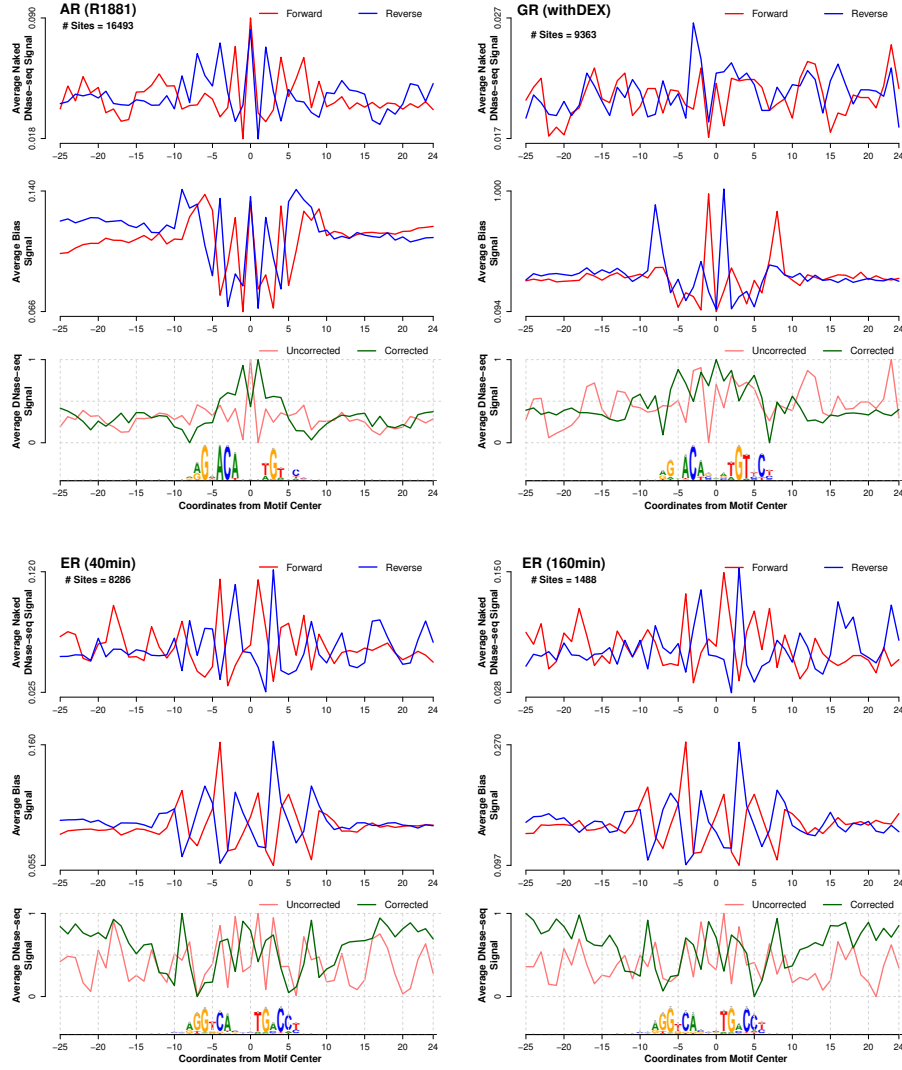


Figure 6: Average DNase-seq signals around nuclear receptor TFs with ChIP-seq evidence in LNCaP(DU), m3134(UW) and MCF-7(DU) cell types. In the top panel, we show the strand-specific average DNase-seq signal on deproteinized DNA experiments (MCF-7(DU) for data sets from single hit and IMR90(UW) for data sets with double-hit protocol); the middle panel shows the strand-specific estimated cleavage bias signal; and the bottom panels shows the (1) uncorrected – observed DNase-seq I cleavage signal and (2) corrected – DNase-seq signal after the bias correction by using Eq. 8. Bottom panel signals were standardized to be in $[0,1]$. Below the graphs, it is shown the motif logo estimated on the DNA sequences of these regions. While corrected DNase-seq profiles from ER have a better match with the underlying motif, this is not the case for AR and GR. However, we observed a small gain in the AUC score comparing HINT-BC and HINT. This difference is in the upper quartile range for all 233 TFs analyzed. These results indicate that cleavage bias correction also brings improvements to footprint prediction of nuclear receptors. However, all these factors have low AUC scores in all footprinting methods, i.e. lower quartiles for HINT-BC or TC AUC score. This indicates that short binding time indeed poses a challenge in footprint prediction.

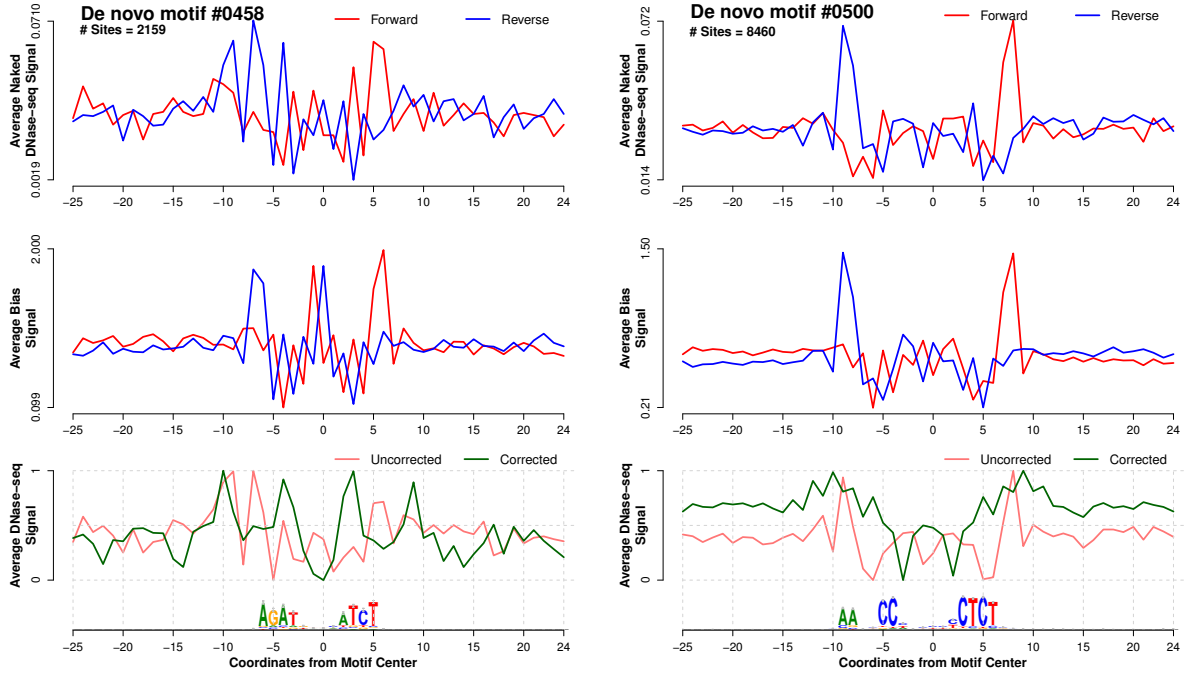


Figure 7: Average bias and DNase-seq signals around binding sites of *de novo* motifs 0458 and 0500 on cell type H7-hESC. In the top panel, we show the strand-specific average DNase-seq signal on deproteinized DNA experiments (MCF-7 cell type); the middle panel shows the strand-specific estimated cleavage bias signal; and the bottom panels shows the (1) uncorrected – observed DNase-seq I cleavage signal and (2) corrected – DNase-seq signal after the bias correction by using Eq. 8. Bottom panel signals were standardized to be in $[0,1]$. Below the graphs, it is shown the motif logo estimated on the DNA sequences of these regions. These motifs were discovered in the footprint analysis of Neph *et al.* (2012) and indicated in He *et al.* (2014) to be artifacts of cleavage bias. Cleavage bias-corrected DNase-seq profiles reveal no clear footprint shape. Furthermore, we compared the overlap between footprints generated by HINT-BC and Neph in H7-hESC(UW) cells. We considered only the MPBSs that overlapped DHSs in H7-hESC. We observed that 24.99% (motif 0458) and 28.58% (motif 0500) of MPBSs associated with a Neph footprint. In contrast, only 0.73% (motif 0458) and 1.71% (motif 0500) of MPBSs overlapped with a HINT-BC footprint. Altogether, this indicates that these motifs are indeed potential artifacts of cleavage bias and reinforces the importance of bias correction prior to any DNase-seq analysis.

Table 1: Summary of DNase-seq data.

Cell Type	Lab	UCSC	GEO/NCBI	# Mapped Reads
H1-hESC	Crawford	wgEncodeEH000556	GSM816632	110303078
HeLa-S3	Crawford	wgEncodeEH000540	GSM816643	54267867
HepG2	Crawford	wgEncodeEH000537	GSM816662	50838536
Huvec	Crawford	wgEncodeEH000548	GSM816646	31848532
K562	Crawford	wgEncodeEH000530	–	365820647
LNCaP	Crawford	wgEncodeEH001097	GSM816637	163625945
MCF-7	Crawford	wgEncodeEH000579	GSM816627	89113893
K562*	Crawford	–	GSM1496625	202001412
MCF-7*	Crawford	–	GSM1496626	210715393
H7-hESC	Stamatoyannopoulos	wgEncodeEH000511	GSM736638 GSM736610	302050785
HepG2	Stamatoyannopoulos	wgEncodeEH000482	GSM736637 GSM736639	168883956
Huvec	Stamatoyannopoulos	wgEncodeEH000488	GSM736575 GSM736533	429088276
K562	Stamatoyannopoulos	wgEncodeEH000484	GSM736629 GSM736566	179970820
m3134	Stamatoyannopoulos	wgEncodeEM001721	GSM1014196	127594903
IMR90*	Stamatoyannopoulos	–	SRA068503	138604440

*Deproteinized DNase-seq experiments.

Table 2: Summary of computational footprinting methods. We characterize distinct methods based on: (1) the approach used for detection of footprints (site-centric or segmentation-based); (2) presence of a smoothing technique on DNase-seq signals; (3) correction of cleavage bias; (4) bias-correction strategy; (5) whether the method is affected by cleavage bias and (6) whether the method significantly outperforms TC. Concerning bias invariance, methods that perform bias-correction of 6-mers (FLR, HINT-BC, HINT-BCN) or that work on smoothed signals (PIQ and Cuellar) do not have their performance influenced by cleavage bias. Note that the DNase2TF implementation only allows the use of 2 or 4-mers and would be possibly improved if 6-mers were supported. Moreover, smoothing of DNase-seq signal as performed by PIQ and Cuellar is also an alternative to implicitly correct for cleavage bias. Concerning prediction performance, all segmentation-based methods (with the exception of Neph) are able to outperform TC prediction performance, while the only site-centric method outperforming TC is PIQ. This indicates an advantage of segmentation-based approaches on the footprint detection problem. Moreover, segmentation-based methods are simpler to execute (single run per DNase-seq experiment) and worked well with default parameters. It is important to point that there is no code available for Boyle method, which makes its usage on further DNase-seq experiments not possible. Neph, Wellington and DNase2Tf methods are based on optimization of flanking regions to estimate FS-like scores. DNase2TF, which is only outperformed by HINT methods, is another good option. DNase2TF did not require extra parametrization and its execution was very straightforward. Site-centric approaches, as PIQ, FLR and Centipede, are estimated for each motif at hand. We have previously observed that Centipede EM-like algorithm has convergence problems for particular factors and required extra parametrization experiments Gusmao *et al.* (2014). This was particularly the case for TF data sets with higher number of MPBSs and high proportion of MPBSs not supported by ChIP-seq (negative examples). A similar behavior is also observed for FLR, which also required the execution of several initializations to avoid numerical problems. Indeed, we could not obtain FLR results for 3 TFs (REST binding on H1-hESC(DU) and K562(DU) and MEF2A binding on H1-hESC(DU)) after executing jobs for 3 weeks. Concerning Cuellar method, its poor predicted performance is due to its simplistic model for the DNase-seq data; average amount of DNase-seq reads on 200 bp reads. PIQ, which has overall good performance, did not show factor-specific issues or required further parametrization experiments. Also, its implementation allows for the execution on several motifs at a time without large computational demand.

Method	Type	Smoothing	Bias Correction	Invariance Bias	Improve TC
Centipede	site-centric				
Neph	segmentation				
FLR	site-centric		6-mers	X	
Cuellar	site-centric	X		X	
Wellington	segmentation				X
Boyle	segmentation				X
DNase2TF	segmentation		4-mers		X
PIQ	site-centric	X		X	X
HINT	segmentation			X	X
HINT-BC	segmentation		6-mers	X	X
HINT-BCN	segmentation		6-mers	X	X

Table 3: Friedman-Nemenyi hypothesis test results on AUC for all evaluated methods. The asterisk and the cross, respectively, indicate that the method in the column outperformed the method in the row with significance levels of 0.01 and 0.05

	HINT-BC	HINT-BCN	HINT	DNase2TF	PIQ	Wellington	Neph	Boyle	FLR	Centipede	Cuellar	TC	PWM	FS
HINT-BC														
HINT-BCN														
HINT	*													
DNase2TF	*	+												
PIQ	*	*	+											
Wellington	*	*	*	*										
Neph	*	*	*	*										
Boyle	*	*	*	*										
FLR	*	*	*	*	*	*	*							
Centipede	*	*	*	*	*	*	*	+						
Cuellar	*	*	*	*	*	*	*	*						
TC	*	*	*	*	*	*	*	*						
PWM	*	*	*	*	*	*	*	*	*	*	+			
FS	*	*	*	*	*	*	*	*	*	*	*	*		