Supplement materials for: Analysis of computational footprinting methods for DNase sequencing experiments

Eduardo G. Gusmao[1,2], Manuel Allhoff[1,2,3], Martin Zenke[1,2], 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, Aachen, Germany.

*ivan.costa@rwth-aachen.de

Supplementary Methods

Supplementary Figures

Supplementary Tables

1 Methods

1.1 Data

DNase-seq aligned reads were obtained from ENCODE (ENCODE Project Consortium, 2012). To perform the computational footprint experiments, 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 cell types H7-hESC, HepG2, Huvec, K562 and m3134 from Stamatoyannopoulous' 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 (Stamatoyannopoulous 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. In addition, to perform the DNase-seq bias estimation clustering, we used all cell types from ENCODE's Tier 1 and Tier 2 cell types, generated in Crawford's and Stamatoyannopoulous' labs. See Supplementary Table 2 for a full DNase-seq 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 the sequence read archive (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. Expression of cells H1-hESC, K562 and GM12878 were obtained from ENCODE (GSE12760 and GSE14863).

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 Datasets 1a and 2b-d.

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

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

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.

DHSs 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 $\mathbf{x}^{\text{fseq}}$ 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

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

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 bias of DNase-seq experiments: (1) aligned reads inside DHSs from proteinized DNase-seq experiments (termed “DNase-seq experimental bias”) following He *et al.* (2014) and (2) all aligned reads for deproteinized (naked) DNA experiments following Yardımcı *et al.* (2014) (termed “DNase-seq cleavage bias”). 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 DNase-seq experimental bias approach.

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

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

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

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

Similarly, the background cleavage score r_w can be evaluated as

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

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

$$b_i^s = \frac{o_w^s \cdot R}{r_w^s \cdot O^s}, \tag{3}$$

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

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

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

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

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 (DNase-seq experimental bias approach); (2) the entire genome (DNase-seq cleavage bias 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

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

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

$$c_i^s = \hat{x}_i^s \hat{b}_i^s. \tag{7}$$

Finally, the bias-corrected DNase-seq genomic signal ($\mathbf{y}$) can be obtained by applying

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

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. Computational resources necessary to the execution of each method were summarized in Supplementary Table 1.

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 (FS) 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 other DNase-seq experiments, 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 an FDR of 1%, which was estimated based on the FS distribution in each cell type (Neph *et al.*, 2012). Finally, we consider only predictions that occurred within DNase-seq hotspots, evaluated using the method first described in Sabo *et al.* (2004). We obtained all hotspots generated by Stamatoyannopoulous' lab in ENCODE for cell types GM12878 (wgEncodeEH000492; GSM736496 and GSM736620), H1-hESC (wgEncodeEH000496; GSM736582) and K562 (wgEncodeEH000484; GSM736629 and GSM736566). 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, which gathers experimental and genomic information around motif-predicted binding sites (MPBSs). It then uses a Bayesian mixture model 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 FDR 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 TFs. We obtained PIQ implementation in `http://piq.csail.mit.edu` and executed it with default parameters, which can be found in the script $c$ommon.r. Briefly, MPBSs were generated with the script $p$wmmatch.exact.r. The DNase-seq signal was created using the script $b$am2rdata.r. And the footprints were detected with the script $p$ertf.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 TFBS (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 I 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: $m$inw $= 6$, $m$axw $= 30$, 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 a single manually annotated genomic region. 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 “DNase-seq experimental bias”, and HINT bias-corrected naked DNA (HINT-BCN) for “DNase-seq cleveage bias” estimated on naked DNA. 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 BinDNase

BinDNase is a site-centric method based on logistic regression to predict active/inactive MBBSs (Kähärä and Lähdesmäki, 2015). The algorithm starts with base pair resolution DNase-seq signal around the MPBSs (+/- 100 bps) and selects discriminatory features using a backward greedy approach. As a supervised approach, the method requires positive and negative examples, which can be obtained from TF ChIP-seq data. We have used DNase-seq data around MPBSs on chromosome 1 for training. These MPBSs were subsequently removed from the evaluation procedure. The definition of positive and negative examples was the same as in our evaluation data sets (Section 1.4.2). Note that this is the only method evaluated here which requires TF ChIP-seq examples for training. We also point the fact that BinDNase did not successfully executed for 19 TFs of our evaluation data set (POU5F1, REST, RFX5, SP1, SP2, SRF, TCF12 and ZNF143 binding in H1-hESC; ARID3A, CTCF, IRF1, MEF2A, PU1, REST, RFX5, SP1, SP2, STAT2 and ZNF263 binding in K562) given our maximum running time criteria (three weeks). Method implementation was obtained at `http://research.ics.aalto.fi/csb/software/bindnase/` and required/provided no parameter selection.

1.3.11 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

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

where $\mathrm{MPBS}_i = [m_i, n_i]$ is the i-th MPBS which extends from genomic positions m_i to n_i and $\overline{\mathrm{MPBS}_i} = (m+n)/2$. 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

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

1.3.12 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

$$\mathrm{TC}_{\mathrm{MPBS}_i} = \sum_{j=\overline{\mathrm{MPBS}_i}-100}^{\overline{\mathrm{MPBS}_i}+99} x_j. \tag{11}$$

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 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 TF ChIP-seq based evaluation

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). Briefly, ROC curves describe the sensitivity (recall) increase as we decrease the specificity of the method. The area under the ROC curve (AUC) metric was evaluated at 1%, 10% and 100% false positive rates (FPR). We also evaluated the area under the precision-recall curve (AUPR). This metric is indicated for problems with imbalanced data sets (distinct number of positive and negative examples) (Davis and Goadrich, 2006; Fawcett, 2006).

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 distinct probability thresholds of (0.8, 0.85, 0.9, 0.95, 0.99) for detection of footprints on all site-centric methods. We performed additional experiments to select the best threshold per method (see Supplementary Fig. 7).

Our TF ChIP-seq based 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 by 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 TF ChIP-seq data sets evaluated in our study: K562(DU) with 59 TFs and H1hesc(DU) with 29 TFs. 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 TFs 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 TF binding time. In this scenario we evaluated the methods PWM, FS, TC, HINT, HINT-BC and HINT-BCN.

Results for TF ChIP-Seq based evaluation are summarized in Supplemmentary Dataset 1.

1.4.3 Expression based evaluation (FLR-Exp)

As shown in Yardımcı *et al.* (2014), ChIP-seq evaluation of putative TFBSs may present biases regarding the fact that ChIP-seq data alone is not able to distinguish direct from indirect binding events. Consequently, we performed an evaluation procedure which combines MPBSs with differentially expressed genes from two cell types. The method evaluates the association of the quality of footprints overllaping particular motifs and the expression of the TF.

We used limma (Ritchie *et al.*, 2015) to perform between-array normalization on expression of H1-hESC, K562 and GM12878 cells and obtain fold change estimates. Then, we retrieved all non-redudant PFMs from Jaspar in which gene symbol is a perfect match with genes present in the array platform. This leads us to 143 PFMs (see Supplementary Datasets 2b–d). We applied a genome-wide motif matching (see Section 1.4.1) using these PFMs.

Afterwards, we evaluated the FLR score (Section 1.3.7), TC (Section 1.3.12) and FS (Section 1.3.11) for the footprints for each of the evaluated method, which intersects with MPBSs of a particular motif. We only considered the footprints within DHSs that are in common between the cell type pair being evaluated, as described in Yardımcı *et al.* (2014). We expect that TFs expressed in cell type A would present higher values regarding these metrics (FLR, TC and FS) with DNase-seq from cell type A in comparisson with these metrics evaluated with DNase-seq from cell type B, and vice-versa. We used a two-sample Kolmogorov-Smirnov (KS) test to assess the difference between each metrics' distribution between the two cell types being evaluated. The KS statistic, which varies from 0 to 1, is used to indicate the difference between two distributions; higher values indicate higher differences. As the KS score do not indicate the direction of the changes in distribution, we obtained a signed version by multiplying KS statistic by -1, in cases where the median of A $<$ median of B. We calculate the Spearman correlation between the signed KS test statistic and the expression fold change for each TF (see Supplementary Fig.2 and 2). Positive values indicate an association between expression of TFs and quality of footprint predictions. We will call this correlation "FLR-Exp". Results for FLR-Exp analysis are summarized in Supplementary Dataset 2a.

1.4.4 Statistical Methods

The non-parametric Friedman-Nemenyi hypothesis test (Demšar, 2006) was used to compare the AUC/AUPR 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. All correlations are based on Spearman values. 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 TFs with shorter binding times. More formally, the protection score for a set of **MPBS** is defined as:

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

where **MBPS** $= \{\text{MPBS}_1, ..., \text{MPBS}_N\}$ is set of binding sites for a given motif, $\text{MPBS}_i = [m_i, n_i]$ is the genomic location of the ith binding site and $n_{C,i}$, $n_{L,i}$ $n_{L,i}$ are the number of DNase-seq 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 indicate no protection. The protection score is a similar statistic as the FS (Sec. 1.3.11). 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 TF 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. Results for protection scores are provided in Supplemmentary Dataset 1.

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.

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA. ACM.

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.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett.*, **27**(8), 861–874.

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.

Kähärä, J. and Lähdesmäki, H. (2015). BinDNase: A discriminatory approach for transcription factor binding prediction using DNase i hypersensitivity data. *Bioinformatics*.

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.

Ritchie, M. E., *et al.* (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, **43**(7), gkv007–e47.

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+.

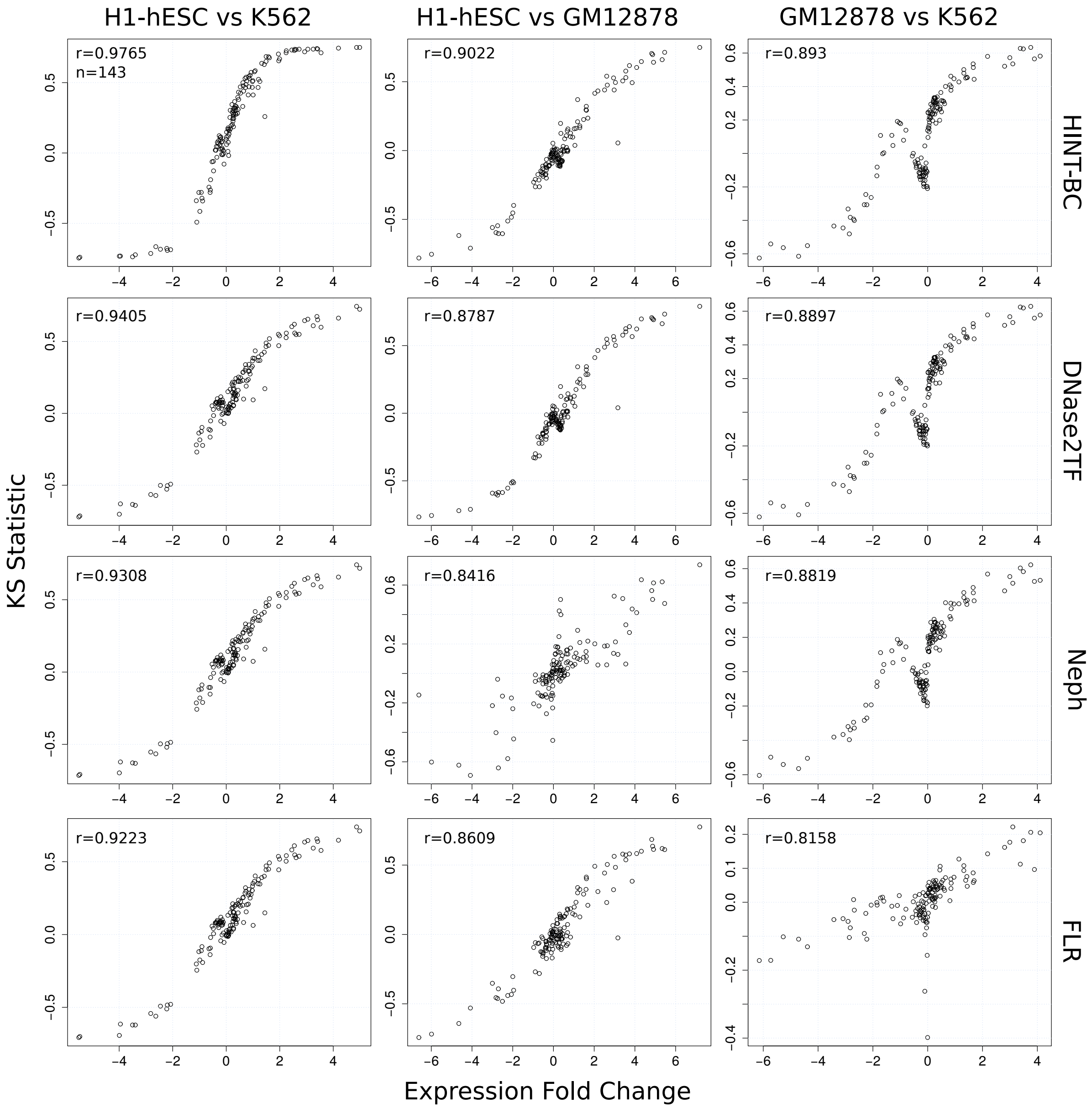

Figure 1: Correlation between KS statistics from FLR scores *vs* fold change expression for cell type pairs H1-hESC *vs* K562 (left), H1-hESC *vs* GM12878 (middle) and GM12878 *vs* K562 (right) for footprints predicted by: HINT-BC, DNase2TF, Neph and FLR (from top to bottom, respectively). We observe high FLR-Exp (Spearman correlation) values (> 0.8) for all cases. Moreover, similar rankings of methods is obtained on the FLR-Exp for each cell pair: H1-hESC/K562 *vs* H1-hESC/GM12878 $r = 0.99$, H1-hESC/K562 *vs* GM12878/K562 $r = 0.96$, and H1-hESC/GM12878 *vs* GM12878/K562 $r = 0.97$.

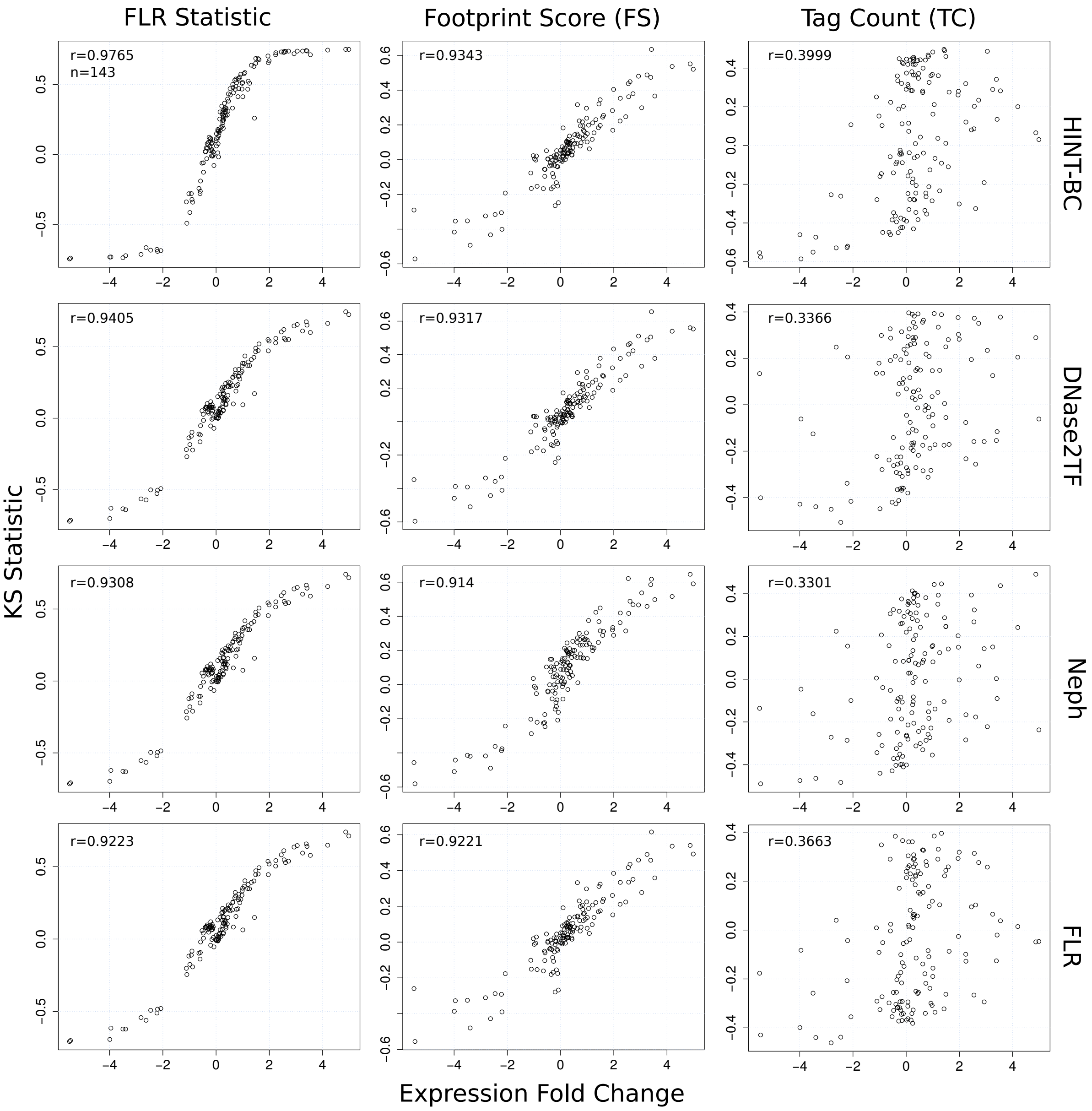

Figure 2: Correlation between KS-statistics *vs* fold change expression for cell type pair H1-hESC *vs* K562 by evaluating either the FLR (left), FS (middle) and TC (right) as quality metric for the footprints. Footprints were predicted with HINT-BC, DNase2TF, Neph and FLR (from top to bottom, respectively). The use of FLR as quality metric presents the highest Spearman correlation values (FLR-Exp). On the other hand, TC exhibits small correlation values (< 0.4) and presents several cases in which the signal of KS and fold change disagree (off diagonal points). Note that the use of FS also have a high average correlation with fold change expression on all evaluated data/methods (average r=0.73) and indicates a ranking of footprint methods similar to FLR ($r = 0.89$). Therefore, it can be used as an alternative metric to FLR-Exp score.

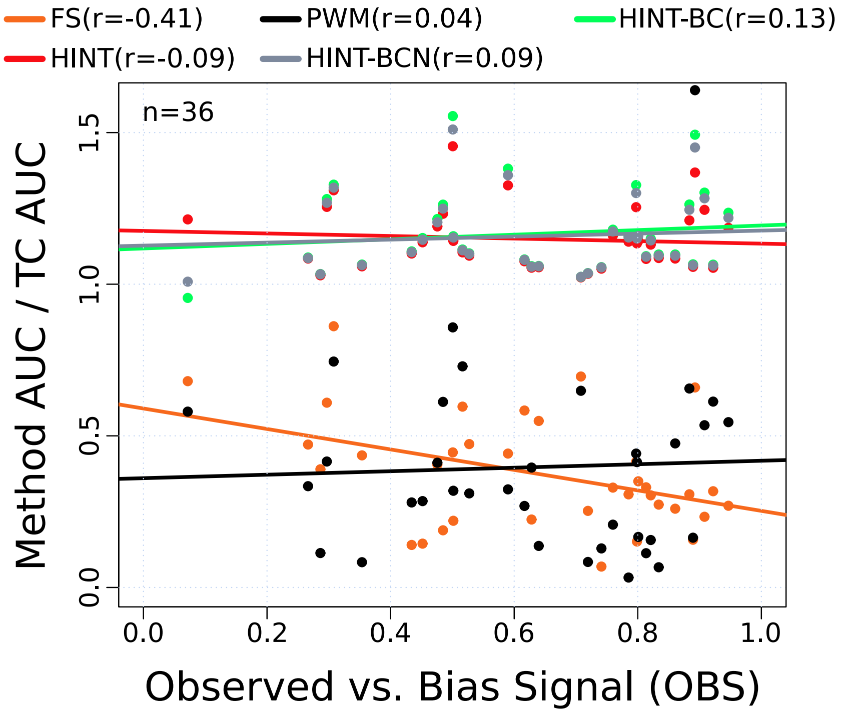

Figure 3: Correlation between the performance of methods and their OBS on `He Dataset`. The x-axis represents the observed sequence bias. The y-axis represents the ratio between the AUC at 10% FPR for a particular method and the TC method. In accordance with He *et al.* (2014), we observe that FS method 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 DHSs 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).

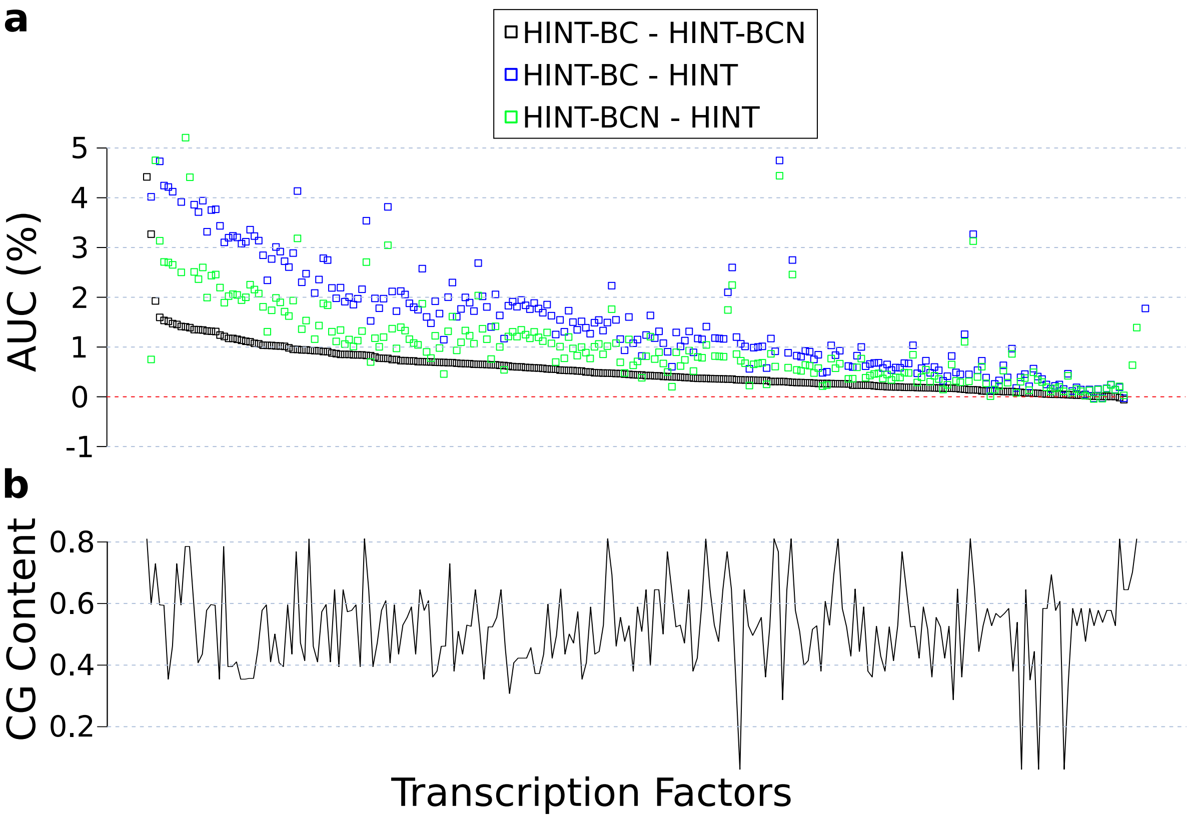

Figure 4: Evaluation of bias correction strategies and CG content contribution. (a) Distribution of AUC (10% FPR) differences bettween HINT-BC and HINT, HINT-BCN and HINT; HINT-BC and HINT-BCN for all 233 TFs of the comprehensive dataset. TFs are ranked by the difference between HINT-BC and HINT-BCN. There is a clear increase in AUC values in the comparison of bias corrected methods (HINT-BC and HINT-BCN) (p-value $< 10^{-30}$; Mann-Whitney-Wilcoxon test). Moreover, HINT-BC has higher AUC values for all but 7 TFs in the comparison with HINT-BCN. (b) CG content of TF motifs. We observe no correlation between CG content of the motifs and the individual AUC of each method: HINT $r = 0.0144$, HINT-BC $r = 0.0254$ and HINT-BCN $r = 0.0108$ (p-value > 0.05; Spearman correlation test). Furthermore, we observe no correlation between CG content of motifs and differences in AUC: HINT-BC $-$ HINT-BCN $r = 0.0188$, HINT-BC $-$ HINT $r = 0.0724$ and HINT-BCN $-$ HINT $r = 0.0644$ (p-value > 0.05; Spearman correlation test).

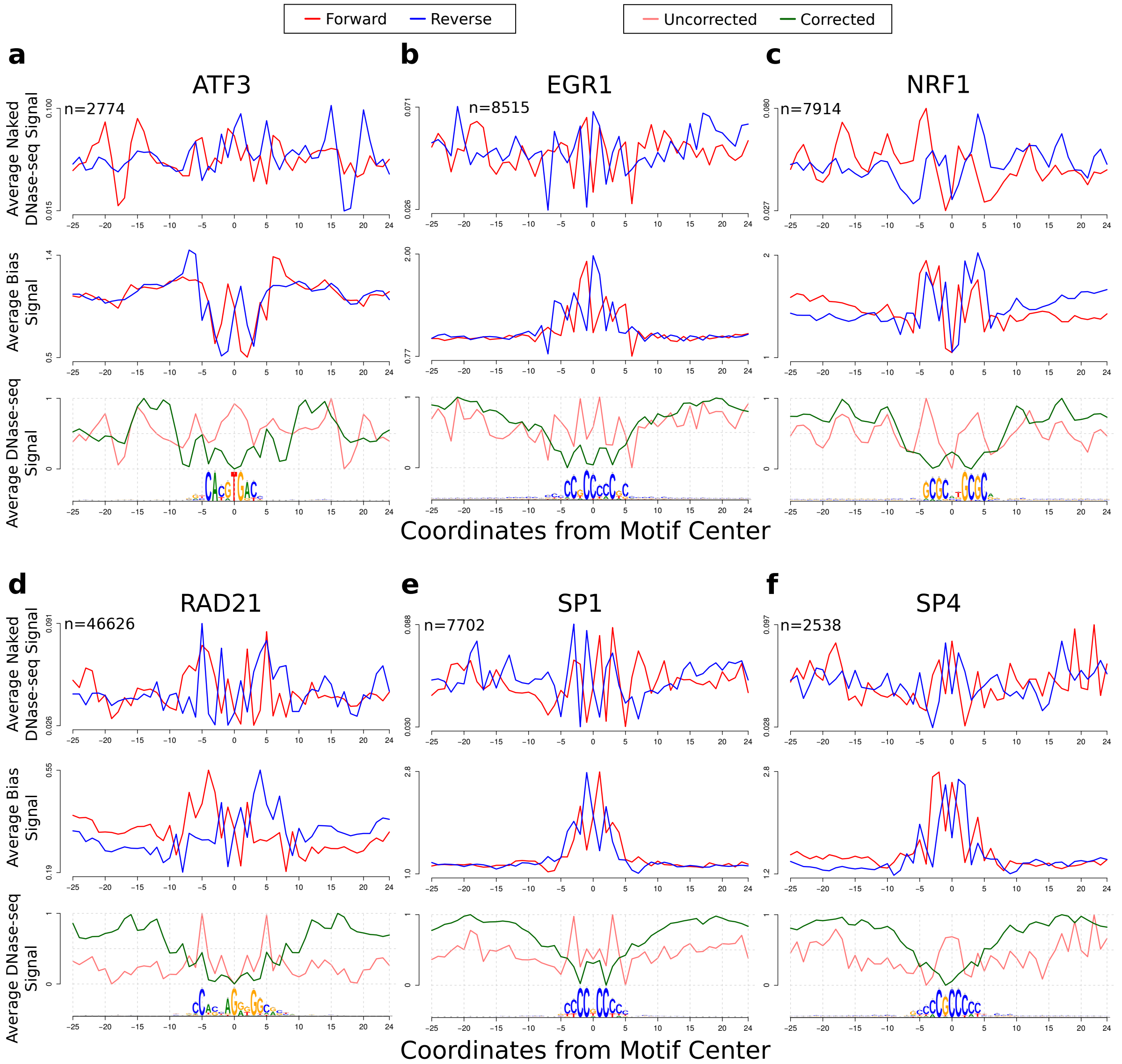

Figure 5: Average DNase-seq signals around selected TFs with ChIP-seq evidence in H1-hESC (DU) cell type. These TFs had the higher AUC gain between HINT-BC and HINT. In the top panel, we show the strand-specific average DNase-seq signal on deproteinized DNA experiments (MCF-7 cell type); the middle panel 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. The bias correction led to a substantial change in the average DNase I cleavage patterns surrounding several TFs. On EGR1, for instance, we observed that the bias-corrected DNase-seq signal presents three clear depletions, which fit the high affinity regions of EGR1 motif (two CC and one C). In contrast, EGR1 uncorrected DNase-seq signal presents a single peak in the center of the motif. The same observations can be made for other TFs, such as NRF1 (with affinity regions (C/G)(C/G)(G/C)C and G(G/C)(C/G)(C/G)C) and SP4 (with affinity region CGCCC). Such patterns reflect bias corrections which are clearly beneficial to footprinting method accuracy.

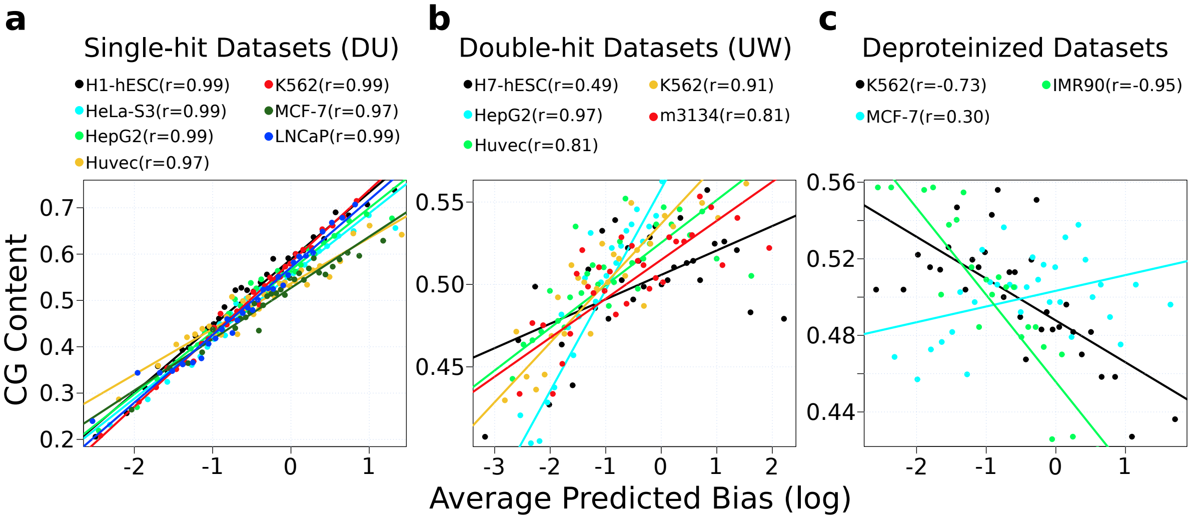

Figure 6: Association between k-mer CG content and DNase-seq experimental bias. We sorted k-mers by their bias estimates and grouped similar ranked k-mers. We show scatter plots with CG content *vs* average cleavage bias for k-mer groups on DHS-based k-mers estimated from (a) single-hit (DU), (b) double-hit (UW) and (c) 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 (p-value < 0.01). Interestingly, we observe a negative correlation for two deproteinized DNA experiments: K562(DU) and IMR90(UW) (p-value $< 10^{-5}$).

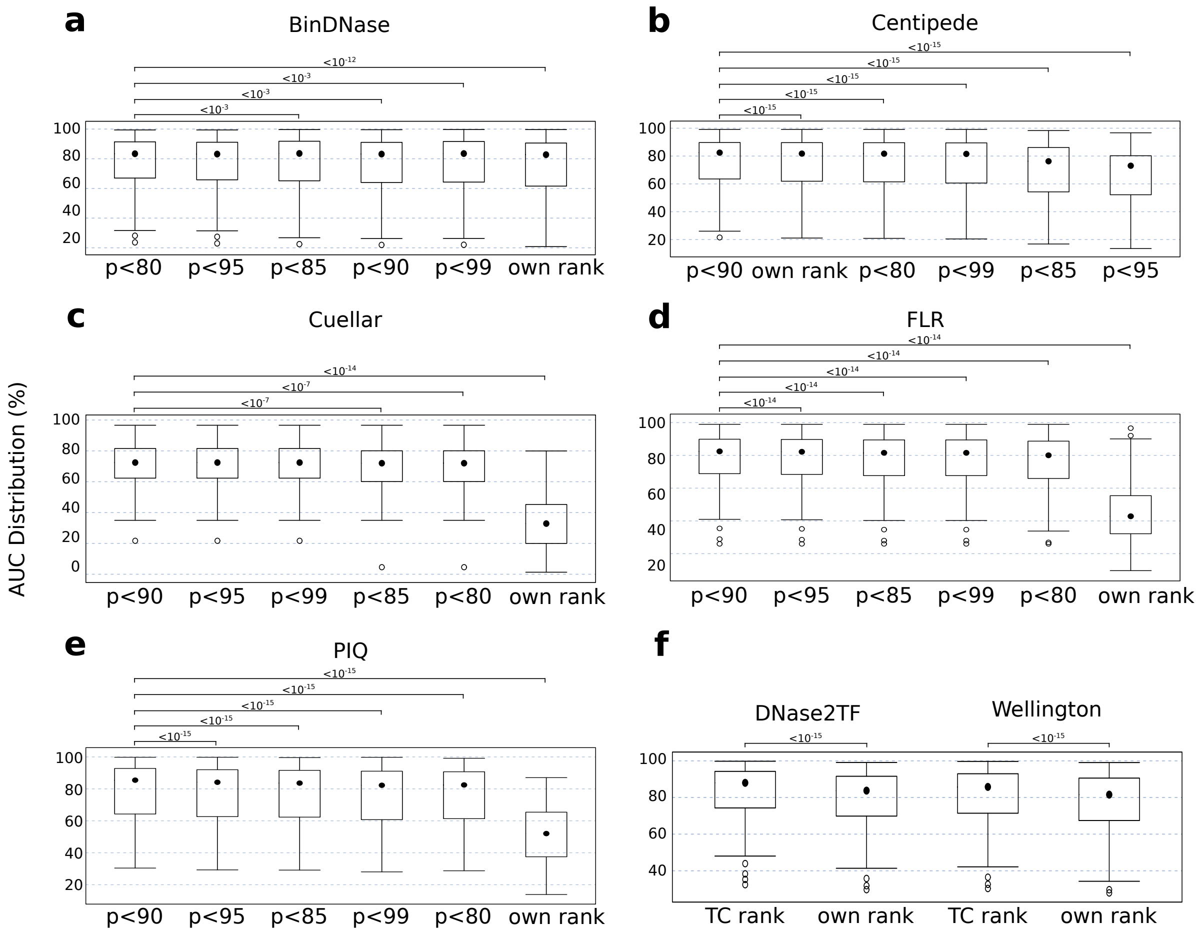

Figure 7: Distribution of AUC values (10% FPR) by using distinct ranking strategies for site centric methods (a) BinDNase, (b) Centipede, (c) Cuellar, (d) FLR, (e) PIQ and (f) segmentation methods DNase2TF and Wellington. Ranking strategies (x-axis) are ranked by decreasing median AUC. In all cases, using TC-based strategies/cutoff was significantly better than the methods original ranking (p-value $< 10^{-12}$; Mann-Whitney-Wilcoxon test). Concerning site-centric methods, the use of a probability threshold of 90% was best for all methods, with the exception of BinDNase, where 80% was best.

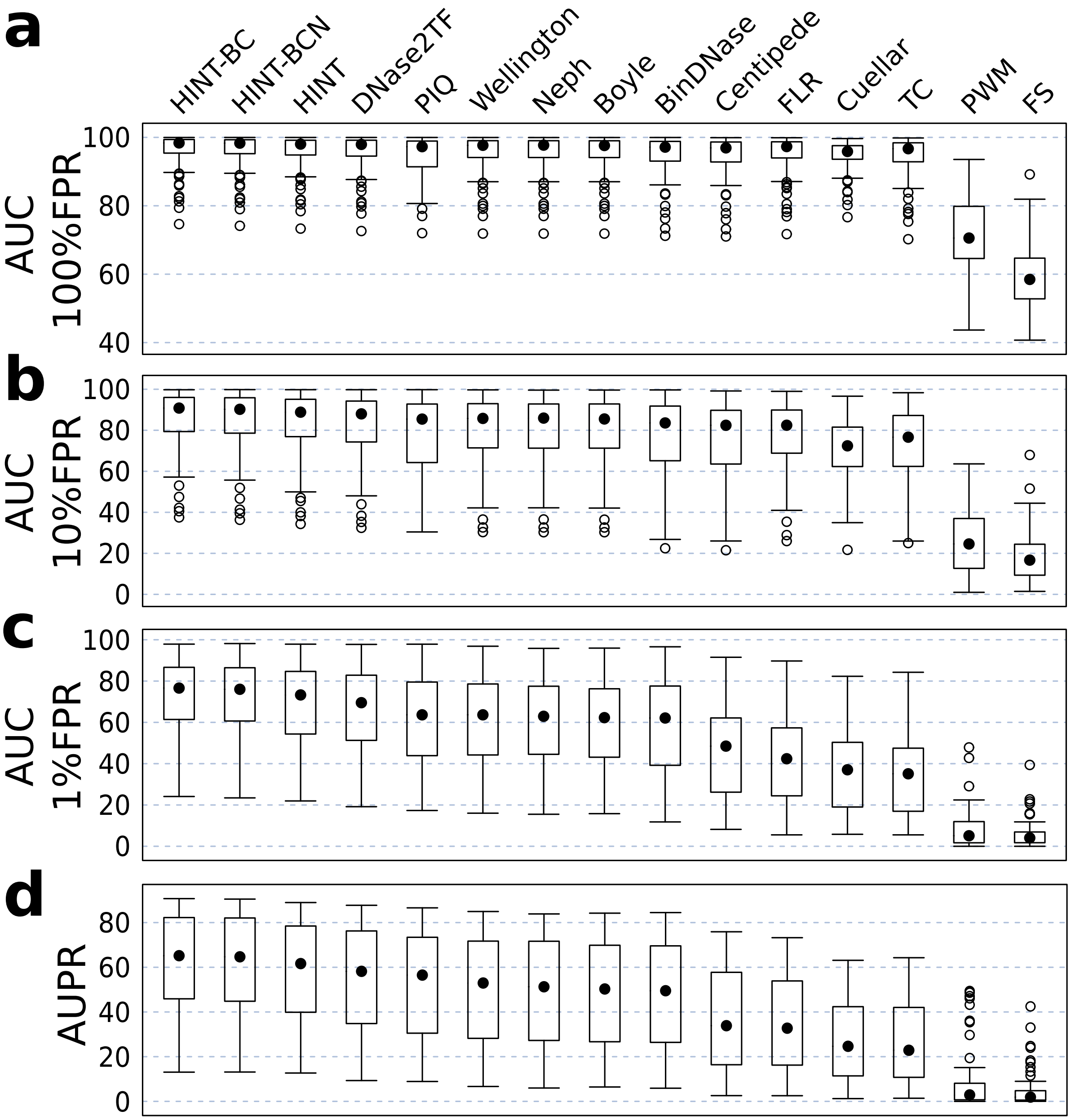

Figure 8: Accuracy distribution for 15 footprinting methods regarding all validation sets (ordered by Friedman Ranking). The accuracy is given by the statistics: (a) AUC at 100% FPR (b) AUC at 10% FPR (c) AUC at 1% FPR and (d) AUPR. We used the Friedman-Nemenyi hypothesis test for statistical evaluation (Demšar, 2006) (Supplementary Tables 3- 6).

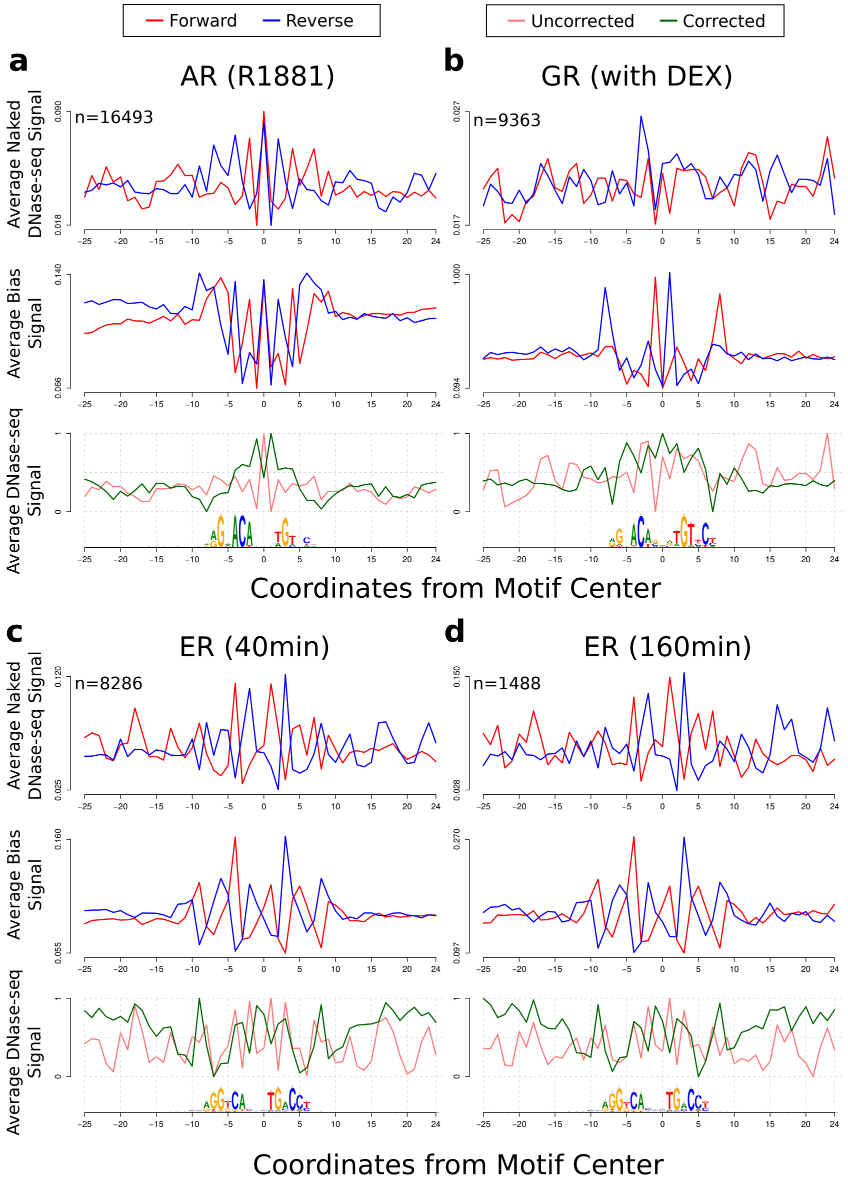

Figure 9: Average DNase-seq signals around nuclear receptor TFs with ChIP-seq evidence in LNCaP(DU), m3134(UW) and MCF-7(DU) cell types. In the top panel, we show the strand-specific average DNase-seq signal on deproteinized DNA experiments (MCF-7(DU) for data sets from single hit and IMR90(UW) for data 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 TFs 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.

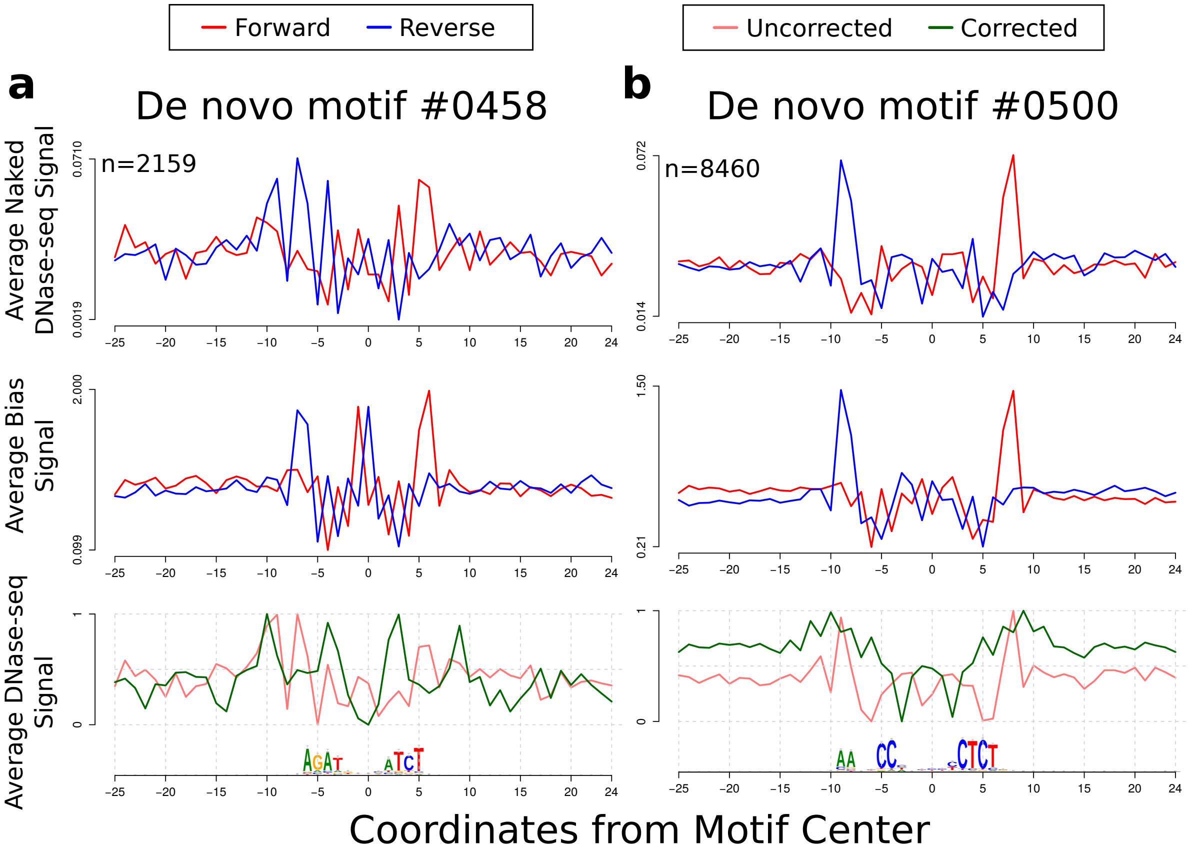

Figure 10: Average bias and DNase-seq signals around binding sites of *de novo* motifs 0458 and 0500 on cell type H7-hESC. In the top panel, we show the strand-specific average DNase-seq signal on deproteinized DNA experiments (MCF-7 cell type); the middle panel shows the strand-specific estimated cleavage 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 computational resources. All methods were run in a Xeon E7-4870 40 CPU (2.4GHz per core). The figures reflect the execution of each method on the `Benchmarking Dataset` (88 TFs; cell types H1-hESC (DU) and K562 (DU)). The table shows the additional steps that the user needs to perform in order to execute the footprinting method, the total CPU time in hours, the maximum memory used during the execution and the total input storage necessary before the execution of each method. Memory comsuption and space requirements of all methods are compatible to a high end desktop. Segmentation based methods, which require a execution per cell, are 4 to 200 times faster than the site-centric methods, which require an execution per cell and TF combination. It is important to notice that PIQ is the only site-centric method, which only requires a single exection per cell. The execution of BinDNAse and Centipede were particularly time consuming (1 week computing on a 40 core server).

Method	Additional Steps	CPU time (hours)	Max. Memory (GB)	Input Storage (GB)
BinDNase	1,2,4	7034	8	95.7
Boyle	NA*	NA*	NA*	NA*
Centipede	1,2,4	7100	8	157.7
Cuellar	1,2,4	575	32	25.4
DNase2TF	3	31	32	29.3
FLR	2,4	870	16	43.1
HINT-BC	3	56	4	17.7
Neph	3	47	4	14.6
PIQ	-	386	32	18.7
Wellington	3	117	16	14.6

[1] Requires extra BAM file input processing.

[2] Requires extra motif matching.

[3] Requires extra DNase-seq peak calling (HS sites).

[4] Requires execution of method for each TF.

[*] Implementation not available.

Table 2: Summary of DNase-seq data. DNase-seq datasets used as input for computational footprinting methods are highlighted in bold. The other DNase-seq datasets were used in the DNase-seq bias estimates clustering analysis. The labs are represented by their aliases: DU (Crawford lab) and UW (Stamatoyannopoulous lab).

Cell Type	Lab	UCSC	GEO/NCBI	# Mapped Reads
H1-hESC	**DU**	**wgEncodeEH000556**	**GSM816632**	**110303078**
HeLa-S3	**DU**	**wgEncodeEH000540**	**GSM816643**	**54267867**
HepG2	**DU**	**wgEncodeEH000537**	**GSM816662**	**50838536**
Huvec	**DU**	**wgEncodeEH000548**	**GSM816646**	**31848532**
K562	**DU**	**wgEncodeEH000530**	**GSM816655**	**365820647**
LNCaP	**DU**	**wgEncodeEH001097**	**GSM816637**	**163625945**
MCF-7	**DU**	**wgEncodeEH000579**	**GSM816627**	**89113893**
K562*	**DU**	–	**GSM1496625**	**202001412**
MCF-7*	**DU**	–	**GSM1496626**	**210715393**
H7-hESC	**UW**	**wgEncodeEH000511**	**GSM736638** **GSM736610**	**302050785**
HepG2	**UW**	**wgEncodeEH000482**	**GSM736637** **GSM736639**	**168883956**
Huvec	**UW**	**wgEncodeEH000488**	**GSM736575** **GSM736533**	**429088276**
K562	**UW**	**wgEncodeEH000484**	**GSM736629** **GSM736566**	**179970820**
m3134	**UW**	**wgEncodeEM001721**	**GSM1014196**	**127594903**
IMR90*	**UW**	–	**SRA068503**	**138604440**
H7hESC	DU	wgEncodeEH002554	GSM1008596	433296955
CD14+	DU	wgEncodeEH003466	GSM1008582	287039145
SKNSH	DU	wgEncodeEH003483	GSM1008585	287186739
MCF7/RandshRNA	DU	wgEncodeEH003468	GSM1008603	288004844
K562/SAHACtrl	DU	wgEncodeEH003489	GSM1008580	503410467
MCF7	DU	wgEncodeEH003470	GSM1008565	89113893
IMR90	DU	wgEncodeEH003482	GSM1008586	303769598
HeLaS3/IFNa4h	DU	wgEncodeEH000577	GSM816633	110348694
K562/G2Mphase	DU	wgEncodeEH003472	GSM1008567	431722812
K562/G1phase	DU	wgEncodeEH003469	GSM1008602	426934260
K562/SAHA1um72h	DU	wgEncodeEH003490	GSM1008558	503301111
MCF7/HypLacAc	DU	wgEncodeEH001745	GSM816670	244207602
K562/NaBut	DU	wgEncodeEH002559	GSM1008601	267722720
CD20+RO01794	DU	wgEncodeEH003465	GSM1008588	256442597
GM12878	DU	wgEncodeEH000534	GSM816665	245090730
A549	DU	wgEncodeEH001095	GSM816649	133567925
MCF7/CTCFshRNA	DU	wgEncodeEH003467	GSM1008581	295954052
K562/ZNFP5	UW	wgEncodeEH003016	–	70400755
CD20+RO01778	UW	wgEncodeEH001884	GSM1024765 GSM1024766	71398619

HeLaS3	UW	wgEncodeEH000495	GSM736510 GSM736564	70669968
K562/ZNF4C50C4	UW	wgEncodeEH003009	–	82579252
A549	UW	wgEncodeEH001180	GSM736506 GSM736580	75764710
K562/ZNFb34A8	UW	wgEncodeEH003012	–	95113482
K562/ZNFg54A11	UW	wgEncodeEH003015	–	76873236
CD14+	UW	wgEncodeEH001196	–	33322702
MCF7/EstCtrl0h	UW	wgEncodeEH003018	GSM1024764 GSM1024767	151170759
MCF7/Est100nm1h	UW	wgEncodeEH003017	GSM1024783 GSM1024784	164440980
K562/ZNF4G7D3	UW	wgEncodeEH003010	–	83034668
K562/ZNFe103C6	UW	wgEncodeEH003013	–	78100065
K562/ZNF2C10C5	UW	wgEncodeEH003008	–	173334712
LHCNM2	UW	wgEncodeEH003005	GSM1024786 GSM1024787	89558026
LHCNM2/Diff4d	UW	wgEncodeEH003006	GSM1024771 GSM1024772	120358720
H1hESC	UW	wgEncodeEH000496	GSM736582	24431583
MCF7	UW	wgEncodeEH000502	GSM736581 GSM736588	89482135
K562/ZNFf41B2	UW	wgEncodeEH003014	–	109124535
CD14+/RO01746	UW	wgEncodeEH001196	GSM1024791	67698560
GM12878	UW	wgEncodeEH000492	GSM736496 GSM736620	47899421
K562/ZNFa41C6	UW	wgEncodeEH003011	–	99106989
HepG2	UW	wgEncodeEH000476	GSM646559	69810990
K562	UW	wgEncodeEH000480	GSM646567	71250291
CD20+RO01778	UW	wgEncodeEH002442	GSM1014525	240594387
K562/ZNFP5	UW	wgEncodeEH003153	–	346226678
K562/ZNFa41C6	UW	wgEncodeEH003152	–	372806338
LHCNM2	UW	wgEncodeEH003149	GSM1014524	255134452
LHCNM2/Diff4d	UW	wgEncodeEH003154	GSM1014539	357827356
H7hESC	UW	wgEncodeEH000834	GSM646563	302050785
HUVEC	UW	wgEncodeEH002460	GSM1014528	429088276
A549	UW	wgEncodeEH003146	GSM1014517	350629033

*Deproteinized DNase-seq experiments.

Table 3: Friedman-Nemenyi hypothesis test results on **AUC at 1% FPR** 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	BinDNase	FLR	Centipede	Cuellar	TC	PWM	FS
HINT-BC															
HINT-BCN															
HINT	*														
DNase2TF	*														
PIQ	*	*													
Wellington	*	*	*	*											
Neph	*	*	*	*	+										
BinDNase	*	*	*	*	*										
Boyle	*	*	*	*	*										
Centipede	*	*	*	*	*	*	*	*							
FLR	*	*	*	*	*	*	*	*	*						
Cuellar	*	*	*	*	*	*	*	*	*						
TC	*	*	*	*	*	*	*	*	*	*					
PWM	*	*	*	*	*	*	*	*	*	*	*				
FS	*	*	*	*	*	*	*	*	*	*	*	+			

Table 4: Friedman-Nemenyi hypothesis test results on **AUC at 10% FPR** 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	BinDNase	FLR	Centipede	Cuellar	TC	PWM	FS
HINT-BC															
HINT-BCN															
HINT	$*$														
DNase2TF	$*$														
PIQ	$*$	$*$	$+$												
Wellington	$*$	$*$	$*$	$*$											
Neph	$*$	$*$	$*$	$*$											
Boyle	$*$	$*$	$*$	$*$											
BinDNase	$*$	$*$	$*$	$*$	$*$										
FLR	$*$	$*$	$*$	$*$	$*$	$*$	$*$								
Centipede	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$							
Cuellar	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$						
TC	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$						
PWM	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$+$			
FS	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$	$*$		

Table 5: Friedman-Nemenyi hypothesis test results on **AUC at 100% FPR** 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	BinDNase	FLR	Centipede	Cuellar	TC	PWM	FS
HINT-BC															
HINT-BCN															
HINT	*														
DNase2TF	*														
PIQ	*	*	*												
Wellington	*	*	*	*											
Neph	*	*	*	*											
Boyle	*	*	*	*											
BinDNase	*	*	*	*	*	*									
FLR	*	*	*	*	*	*	*								
Cuellar	*	*	*	*	*	*	*	+							
Centipede	*	*	*	*	*	*	*	*							
TC	*	*	*	*	*	*	*	*	*						
PWM	*	*	*	*	*	*	*	*	*	*	*	*			
FS	*	*	*	*	*	*	*	*	*	*	*	*	*		

Table 6: Friedman-Nemenyi hypothesis test results on **AUPR** 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	BinDNase	FLR	Centipede	Cuellar	TC	PWM	FS
HINT-BC															
HINT-BCN															
HINT	*														
DNase2TF	*	*													
PIQ	*	*													
Wellington	*	*	*												
Neph	*	*	*	*	+										
Boyle	*	*	*	*	*										
BinDNase	*	*	*	*	*										
Centipede	*	*	*	*	*	*	*	+	+						
FLR	*	*	*	*	*	*	*	*	*						
Cuellar	*	*	*	*	*	*	*	*	*	*					
TC	*	*	*	*	*	*	*	*	*	*					
PWM	*	*	*	*	*	*	*	*	*	*	*				
FS	*	*	*	*	*	*	*	*	*	*	*	+			