Correction of DNase-seq cleavage bias impacts on quality of footprinting

Eduardo G. Gusmao[1,2], 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.

1 Introduction

Next-generation sequencing (NGS) technologies coupled with novel or modified biochemical methods enhanced significantly our understanding on diverse molecular biology questions such as the regulatory landscape and chromatin structure/dynamics (Meyer and Liu, 2014). DNase-seq, a technique in which digested chromatin fragments are sequenced to identify accessible regulatory regions, is an example of such NGS-based techniques (Crawford *et al.*, 2006; Song and Crawford, 2010). The nucleotide-level resolution of DNase-seq allows the accurate identification of transcription factor binding sites (TFBSs) through a well-characterized pattern of DNase I cleavage (Hesselberth *et al.*, 2009; Neph *et al.*, 2012; Wang *et al.*, 2012; Natarajan *et al.*, 2012; ENCODE Project Consortium, 2012). Generally, TFBSs present high DNase I digestion levels at their flanking regions and low DNase I digestion levels at their protein-bound positions (footprints) (Boyle *et al.*, 2011).

Computational footprinting methods were developed to identify active TFBSs on a genome-wide scale. Footprinting methods are based on the application of mathematical models on DNase I cleavage profiles. They detect short regions with low DNase I digestion activity (indicative of transcription factor binding) around regions with high digestion activity (naked DNA around TFBS). These methods can be categorized as either segmentation-based or site-centric. Site-centric footprint methods evaluate the DNase-seq signal around motif-predicted binding sites (MPBSs) to classify them as active or inactive (Cuellar-Partida *et al.*, 2012; Pique-Regi *et al.*, 2011; Sherwood *et al.*, 2014; Yardmc *et al.*, 2014). While these methods use powerful discriminative learning frameworks, they depend on an initial sequence-based motif match and

*ivan.costa@rwth-aachen.de

will miss *de novo* motifs. Alternatively, segmentation-based methods evaluate all possible genomic positions to detect footprints only from the DNase-seq signal (Boyle *et al.*, 2011; Gusmao *et al.*, 2014; Neph *et al.*, 2012; Piper *et al.*, 2013; Sung *et al.*, 2014).

It is well known that NGS-based methods for chromatin biology present biases that may result on misleading interpretations (Meyer and Liu, 2014). These biases may arise in many levels of the experimental protocols: chromatin fragmentation and size selection (enzymatic or by sonication), DNA isolation, individual transcription factor (TF) binding characteristics and others (Meyer and Liu, 2014). Of particular interest here is the bias resulting from the DNase I enzyme preference to cleave certain nucleotide sequences. Recently, it was shown that such bias affects DNase-seq cleavage patterns around binding sites of particular TFs (He *et al.*, 2014). Authors compare two simple approaches to classify an MPBS as active or inactive: (1) tag count score (TC), which consists of ranking MPBSs by the number of DNase-seq reads around their binding sites; and (2) a version of the footprint score (FS) proposed in Neph *et al.* (2012), which consists of ranking MPBSs by the ratios of DNase-seq reads within the binding site and in the upstream/downstream flanking regions. They show that the TC approach is better than the FS and that the FS performance deteriorates for TFs with motifs associated to high DNase-seq cleavage bias. Concurrently, Sung *et al.* (2014) indicated that footprints of transcription factors with short binding sites are artifacts of DNase cleavage bias. Authors indicate that factors with short binding time have characteristic low protection footprinting shapes, i.e. low number of DNase I cuts around the footprint. Altogether, these findings question the feasibility of footprinting methods. However, He *et al.* (2014) did not evaluated state-of-the-art computational footprinting methods (Boyle *et al.*, 2011; Cuellar-Partida *et al.*, 2012; Gusmao *et al.*, 2014; Pique-Regi *et al.*, 2011; Piper *et al.*, 2013; Sherwood *et al.*, 2014; Sung *et al.*, 2014; Whitington *et al.*, 2009; Yardmc *et al.*, 2014), or evaluated the effects of correcting the DNase-seq signal prior to the application of footprinting. Moreover, no work has explored the use of footprint protection concept introduced in Sung *et al.* (2014) for computational detection of short time binding factors.

Here, we analyze the effects of the DNase-seq cleavage bias and TFs with short binding residence time (short-lived TFs) on methods previously analyzed by Gusmao *et al.* (2014). Moreover, we use a method to correct DNase-seq cleavage bias in combination with our HMM-based methodology (HINT) (Gusmao *et al.*, 2014). We evaluate the effect of bias correction on a large DNase-seq and ChIP-seq compendia with 13 cells, 144 factors and 13 methods. We show that few footprinting methods are influenced by cleavage bias, which is not the case for methods with a good performance in our previous study. Moreover, the performance of HINT is significantly improved after cleavage bias correction. Lastly, we propose a statistic metric to score the DNase I protection level of TFBSs, which is able to indicate potential short-lived TFs. Moreover, this metric correlates with the accuracy of footprinting methods.

2 Material and Methods

2.1 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 Lab (labeled with the initials of their institution "DU") and concerning 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) (Yardmc *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. See Table 1 for data description.

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 ChIP-seq raw sequences for LNCaP cell type was obtained in Gene Expression Omnibus (GEO) with accession number GSM353644 (Yu *et al.*, 2010). ER ChIP-seq raw sequences for MCF-7 cell type was obtained in GEO with accession number GSE54855 (Guertin *et al.*, 2014). GR 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).

2.2 Bias Correction

2.2.1 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 resulting 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 $s \in \{+, -\}$ by considering the strand the read was mapped to.

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	Stamatoyannopoulous	wgEncodeEH000511	GSM736638 GSM736610	302050785
HepG2	Stamatoyannopoulous	wgEncodeEH000482	GSM736637 GSM736639	168883956
Huvec	Stamatoyannopoulous	wgEncodeEH000488	GSM736575 GSM736533	429088276
K562	Stamatoyannopoulous	wgEncodeEH000484	GSM736629 GSM736566	179970820
m3134	Stamatoyannopoulous	wgEncodeEM001721	GSM1014196	127594903
IMR90*	Stamatoyannopoulous	–	SRA068503	138604440

*Deproteinized DNase-seq experiments.

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 $\mathbf{x}^{\text{fseq}}$ 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].

2.2.2 Estimation of DNase I Cleavage Bias

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

We use two approaches for estimation of the intrinsic DNase I cleavage bias: (1) aligned reads inside DHSs from DNase-seq experiments and (2) all aligned reads for deproteinized (naked) DNA experiments.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. Bias estimation mathematical formalizations will be made based on the DHS-based 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

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

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 $k+1 \leq i \leq N-k+1$, strand s given that $w = G^s[i - \frac{k}{2}..i + \frac{k}{2}]$ can be calculated as

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

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 (DHS approach); (2) the entire genome (deproteinized 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.

2.2.3 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

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

2.3 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 Table 4 shows a summary of all methods evaluated in this study according to many key features from footprinting methods.

2.3.1 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".

2.3.2 Boyle Method

Boyle *et al.* (2011) designed a segmentation-based approach, which is based on using 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/`.

2.3.3 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, 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).

2.3.4 Cuellar Method

Cuellar-Partida *et al.* (2012) proposed a site-centric method to include DNase-seq data as priors for the detection of active 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".

2.3.5 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, Wellignton only considers reads mapped to the 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 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.

We also point that, although Wellington was designed to be executed on DNase-seq data generated using the double-hit protocol, in our experiments it produces similarly good results on data from single-hit protocol.

2.3.6 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 *c*ommon.r. Briefly, putative binding sites 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.

2.3.7 Footprint Mixture (FLR)

Yardmc *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 multinomials, representing active (footprints) and inactive (background) MPBSs. The background model is initialized with either bias cleavage frequencies or estimated *d*e 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 referr to this method as "FLR".

2.3.8 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 interactivelly 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 run DNase2TF with a 4-mer cleavage bias correction. Other parameters were set to their default values: $minw = 6$, $maxw = 30$, $z\_threshold = -2$ and $FDR = 10^{-3}$.

2.3.9 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. 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 pipeline of this method starts by normalizing the DNase I cleavage signal using a global and local 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 predicted 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 also used bias-corrected DNase-seq signal prior to normalization steps. We will call the method HINT bias-corrected (HINT-BC), for correction based on reads inside DHS regions, and HINT bias-corrected naked DNA (HINT-BCN), for bias correction based on DNase-seq experiments on deproteinized DNA. This novel method, scripts for computation of cleavage bias and pre-computed cleavage bias files are available at `www.regulatory-genomics.org/hint`.

2.3.10 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 statistics 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

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

2.3.11 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

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

Scripts for computation of Tag Count and Footprint Score are available at `www.regulatory-genomics.org/hint`.

2.4 Evaluation

2.4.1 Motif-Predicted Binding Sites

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 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".

2.4.2 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 p-value cutoff of 0.9 for all methods. We observed that this approach resulted in significantly higher accuracy for these methods (see Fig. XXX).

Finally, a Friedman-Nemenyi hypothesis test (Demšar, 2006) was used to compare the AUC at the 10% (FPR) 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.

Our comparative experiments comprise the following three evaluation scenarios. All evaluation statistics and method performances are available at the Supplementart File *SupplementaryDataTable.xls*.

`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 H1hesc(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 209 combinations of cells and transcription factors with matching cellular background. This data

set was build 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. We evaluated here the methods PWM, FS, TC, HINT, HINT-BC and HINT-BCN.

2.4.3 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 previsouly 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:

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

where $\mathbf{MBPS} = \{\text{MPBS}_1, ..., \text{MPBS}_N\}$ 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 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. 2.3.10). 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.

3 Results

3.1 DNase I Cleavage Correlation between Data Sets

First, we compared the DNase I cleavage bias correlation in different DNase-seq data sets. For that, we applied a pairwise **Spearman** correlation test for each 6-mer w ratio between observed and expected cleavage bias ($o_w^s \cdot R^s / r_w^s \cdot O^s$).

Fig. 1 exhibits the Ward's minimum variance clustering of correlation coefficient R for all data sets. 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 MCF7(DU)) and group 3 contains K563 and HepG2 from UW. Correlation within all

experiemnts in a group are significant (p-value < 0.001; after Bonferroni correction). These results confirms that experiments from the same lab/protocol have similar bias. The only exceptions were K562(UW) and Hepg2(UW), which presented low correlation values with any other UW and DU experiments. Interestingly, deproteinized DNA experiments from distinct protocols clustered together. The high correlations between deproteinized DNA experiments ($R = 0.94$ for K562 and MCF7; $R = 0.81$ for K562 and IMR90) are in agreement with previous reports Yardmc *et al.* (2014). In particular, deproteinized DNA experiments have a high average correlation with experimets with 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). These results corroborates with the hypothesis that deproteinized DNA could be used for estimation bias correction for experiments with same protocol. A cautionary remark is the lack of correlation between deproteinized IMR90(UW) and K562/HepG2(UW) experiments.

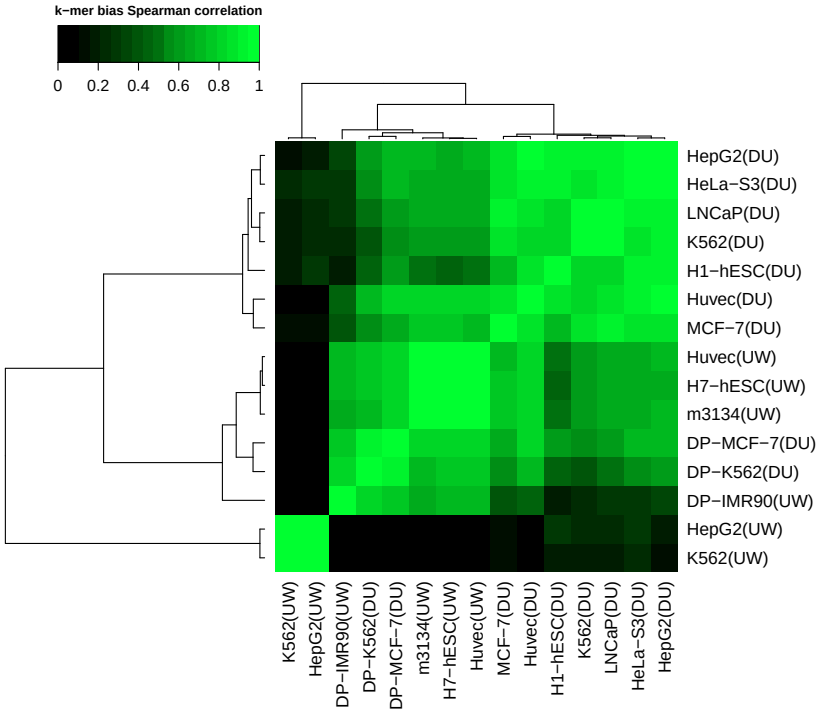

Figure 1: Pairwise Spearman correlation coefficient (R) between different DNase-seq data sets calculated based on each 6-mer w ratio between observed and expected cleavage bias ($o^s_w \cdot R/r^s_w \cdot O^s$). Crawford's lab data sets are represented by "DU" (Duke University; single-hit protocol) and Stamatoyannopoulous' lab data sets are represented by "UW" (University of Washington; double-hit protocol).

3.2 Bias Correction Improves Computational Footprinting Performance

3.2.1 Replication of He et al.

It was shown in He *et al.* (2014) that, compared to TC method, the performance of FS was highly impacted by DNase I cleavage bias. For this, the authors evaluated the correlation between the AUC ratio of FS and TC, and the observed *vs.* predicted bias signals (OBS). OBS is defined as the Pearson correlation between observed DNase I cleavage signal and predicted bias signal evaluated with Eq. 3 around a 50 bp window centered on MPBSs with ChIP-seq evidence. A high OBS means that the predicted bias signal is very similar to the observed cleavage.

We replicated the analysis with the same 36 TFs from He *et al.* (2014) including a PWM approach and digital genomic footprinting methods HINT and HINT-BC (Fig 2). 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 correllation is found for 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 hypersentitivity 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).

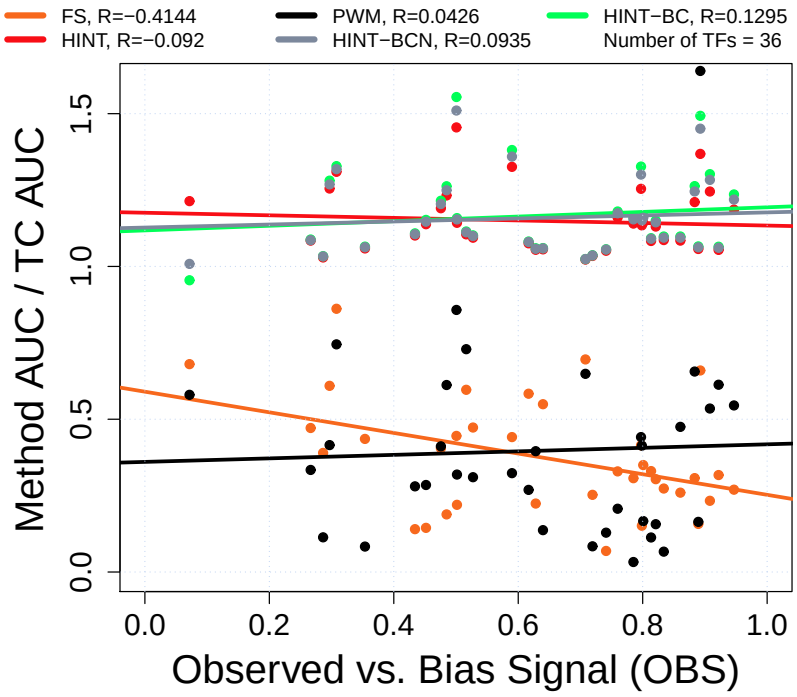

Figure 2: Correlation between the performance of methods and their OBS on `He Dataset`. The x-axis represents the observed sequence bias. The y-axis represents the ratio between the AUC at 10% FPR for a particular method and the TC method.

3.2.2 Influence of Bias in State of Art Methods

To investigate if the influence of DNase-seq cleavage bias on start-of-the-art footprinting methods, we evaluate 14 methods on the `Benchmarking Dataset` with 2 DNase-seq experiments in H1-hESC(DU) and K562(DU) cells and 88 TFs. The Fig. 3 shows the association between prediction performance (using TC performance as reference) and cleavage bias in the benchmarking datasets. The prediction performance of six methods (Wellington, Neph, Boyle, DNase2TF, Centipede and FS) have a significant negative correlation with the observed sequence bias (p-value < 0.05 after Bonferroni correction). However, even though applied to uncorrected DNase-seq signal, other methods (HINT, Cuellar, PIQ and PWM) did not present a significant correlation. Moreover, all methods using 6-mers estimates of cleavage bias for footprint detection (HINT-BC, HINT-BCN and FLR), do not have a significant correlaiton with cleavage bias. In particular, HINT-BC have lowest absolute correlation overall.

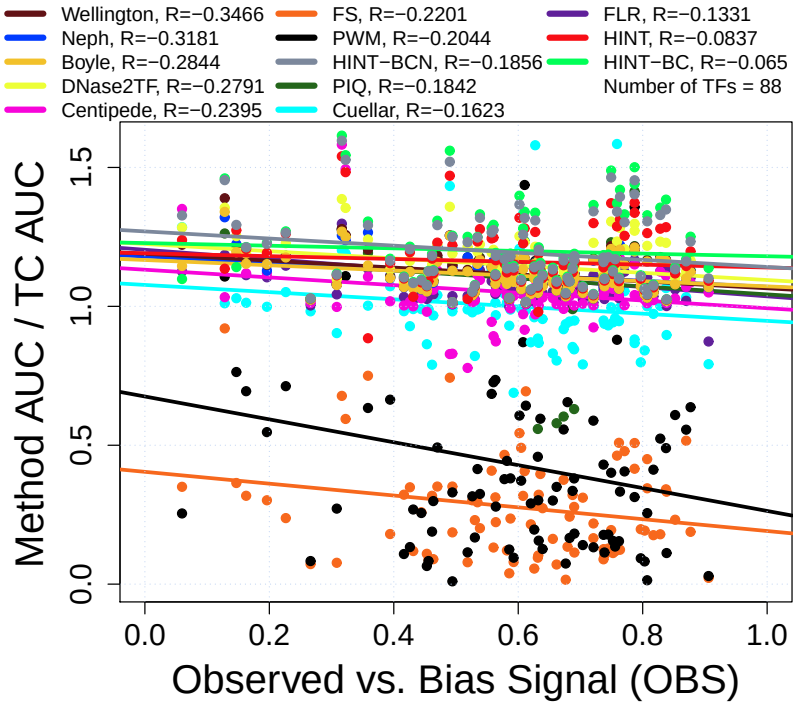

Figure 3: Correlation between the performance of methods and their OBS on `Benchmarking Dataset`. The x-axis represents the observed sequence bias. The y-axis represents the ratio between the AUC at 10% FPR for a particular method and the TC scoring method.

Next, we compared the performance of all methods regarding their AUC (Fig. 4). We used the Friedman-Nemenyi ranking and hypothesis test for statistical evaluation (Demšar, 2006) (Tables 3.2.2 and 3.2.2). 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 < 0.05). Moreover, HINT-BC and HINT-BCN, outperformed all other competing methods. The AUC of the TC approach was significantly higher than FS method (adjusted p-value < 0.05).

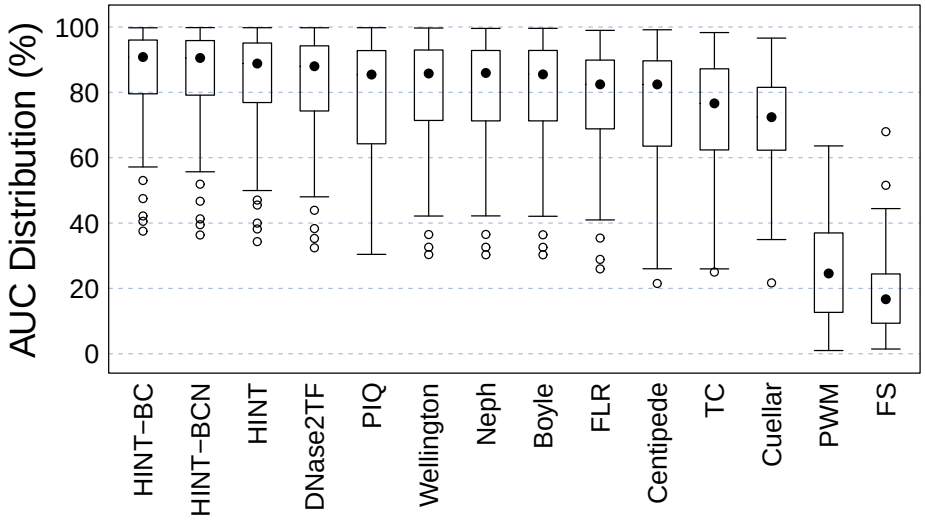

Figure 4: AUC distribution for eight compared methods regarding all validation sets (ordered by median AUC).

Table 2: Friedman ranking. For each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

AUC	
HINT-BC	1.3409
HINT-BCN	2.0682
HINT	3.6023
DNase2TF	4.0682
PIQ	5.6932
Wellington	6.3068
Neph	6.9091
Boyle	7.6477
FLR	9.3807
Centipede	9.7045
Cuellar	10.6761
TC	11.1705
PWM	12.7273
FS	13.7045

We present in Table 4 a summary of characteristics of evaluated method and their outcome on our evaluations. We characterize distinct methods based on: (1) the approach used for detection of footprints (site-centric or segmentation-based); (2) presence of a smoothing on

Table 3: Friedman-Nemenyi hypothesis test results for the AUC metric. The asterisk and the cross, respectively, indicate that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1

	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	*	*	*	*	*	*	*	*	*	*	*	*		

DNase-seq signals and (3) correction of cleavage bias. 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 is also an alternative as it seems to assuage cleavage bias implicitly.

Concerning prediction peformance, 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, these methods are simpler to execute (single run per DNase-seq experiment) and work well on 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.

Models for site-centric approaches, as PIQ, FLR and Centipede, are estimated for each motif at hand. We have previously observerd that Centipede EM-like algoritmh had 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 non supported by ChIP-seq (negative examples). A similar behavior is also observed for FLR, which also required the execution of several initializations to avoid frequent spurious solutions. Indeed, both Centipede and FLR require an execution for each motif of interested imposing large computational and techinical demand. An exceptional site-centric method is PIQ, which have overall good performance, did not showed factor-specific issues or required further parametrization experiments. Also, it can be executed on several motifs at a time and without a large computational demand.

Table 4: Summary of computational footprinting methods.

Method	Type	Smothing	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

3.3 Examples of Uncorrected *vs.* Bias-Corrected DNase-seq Profiles

The bias correction led to a substantial change in the average DNase I cleavage patterns surrounding several TFs. Fig. 5 shows examples of such changes for 6 TFs in cell type H1-hESC (DU) with higher AUC gain between HINT-BC and HINT. 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.

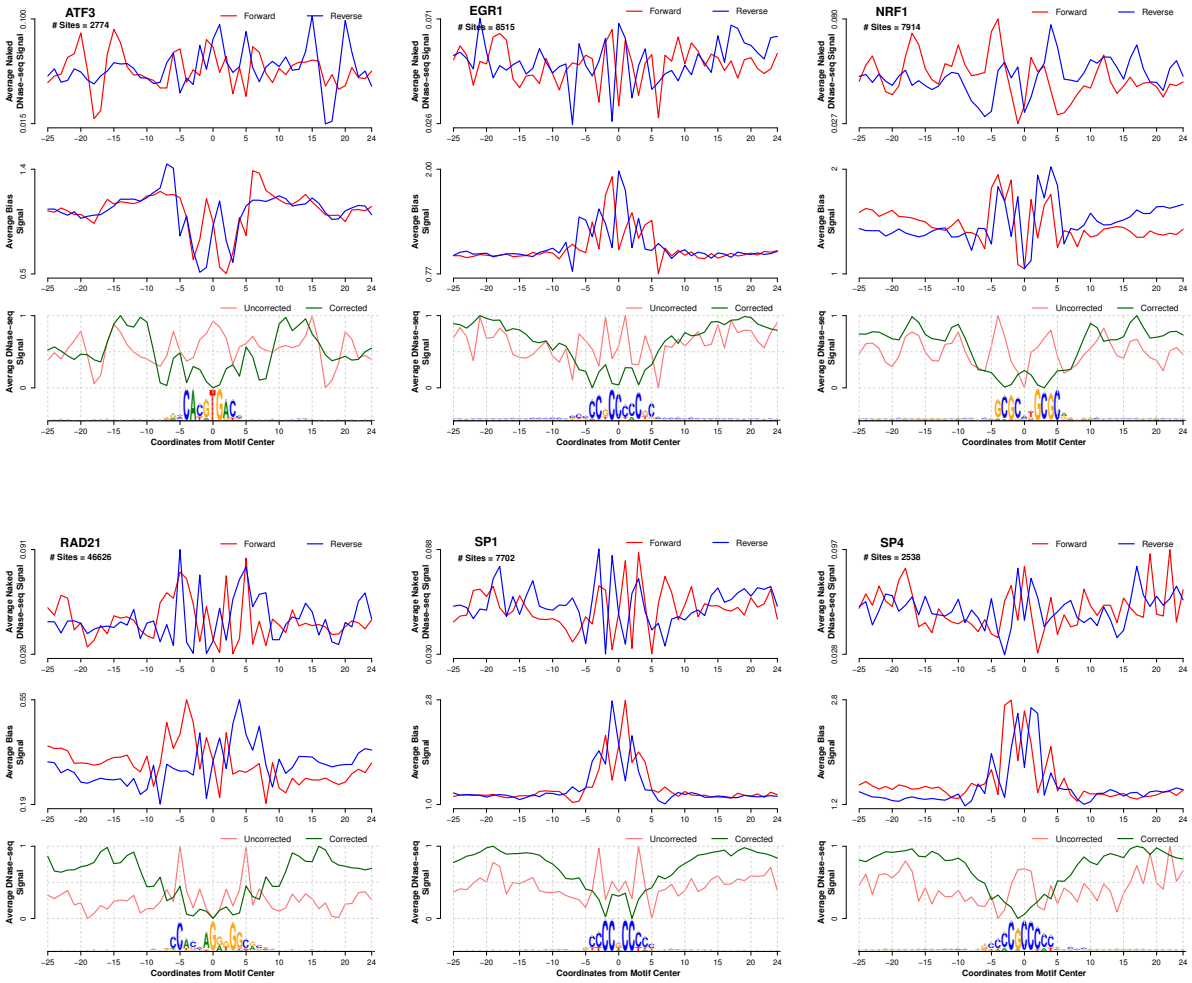

Figure 5: Average DNase-seq signals around selected TFs with ChIP-seq evidence in H1-hESC (DU) cell type. In the top panel, we show the strand-specific average DNase-seq signal on deproteinized DNA experiments (MCF-7 cell type); the middle panel shows the strand-specific estimated cleavage bias signal; 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.

A recent study demonstrated that transcription factors with short binding time have DNase-seq profiles which resemble estimated DNase-seq cleavage bias (Sung *et al.*, 2014). Authors indicate nuclear receptor factors ER and GR as typical exmaples of short time binding factors and AP-1(JUN) and CTCF as example of medium/long binding time factors. He *et al.* (2014) used nuclear receptor factors (AR and GR) as examples of factors with DNase-seq profiles resembling cleavage bias. As shown in Figure 6, bias-corrected DNase-seq average profiles of AR, ER and GR are distinct from either deproteinized DNA and cleveave bias estimates. 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 209 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.

To further investigate the footprinting issue regarding binding time of transcription factors, we propose a statistic called protection score. The protection score measures the average differece in the number of DNase-seq reads in the flanking and binding site regions for a given TF (Section 2.4.3). As indicated in Figure 7, the protection score is negative at binding sites of hormone receptors AR (-0.71), ER (-0.52) and GR (-1.19), i.e. there are more DNase-seq cuts in the binding sites than in the flanking regions at average. Moreover, the protection score can successfully separate factors with short binding time (such as nuclear receptors) from factors with high binding time (AP-1/JUN and CTCF). Fig 7 also reveals a clear association of the proteciton score and the AUC of HINT-BC, i.e. factors with negative protection score have much lower AUCs than factors with positive scores. Overall, the protection score is positively correlated with the AUC values of evaluated methods, i.e. 0.26 with TC AUC (p-value $< 10^{-4}$) and 0.28 with HINT-BC AUC (p-value $< 10^{-4}$), and negativelly correlated with cleavage bias score (-0.44; p-value $< 10^{-4}$). Altogether, this indicates that transitivity of TFs (as measured by the protection score) is a more relevant feature than DNase-seq intrinsic cleavage bias for indicating the performance/feasibility of computational footprinting methods.

Finally, we investigate the profiles of *de novo* motifs 0458 and 0500 motifs discovered in the footprint analysis of Neph *et al.* (2012) and indicated in He *et al.* (2014) to be artifacts of cleavage bias. As shown in figure 8, cleavage corrected DNase-seq profiles reveal no clear fooptrint shape. Moreover, their binding sites have negative protection scores (-1.33 and -0.22 respectivelly). Furthermore, we compared the overlap between footprints generated by HINT-BC/Neph in the cell type in which these factors were identified (H7-hESC(UW)) and MPBSs based on the PWMs provided in Neph *et al.* (2012). 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 previously to any DNase-seq analysis.

References

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.

Crawford, G. E., *et al.* (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, **16**(1), 123–131.

Cuellar-Partida, G., *et al.* (2012). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, **28**(1), 56–62.

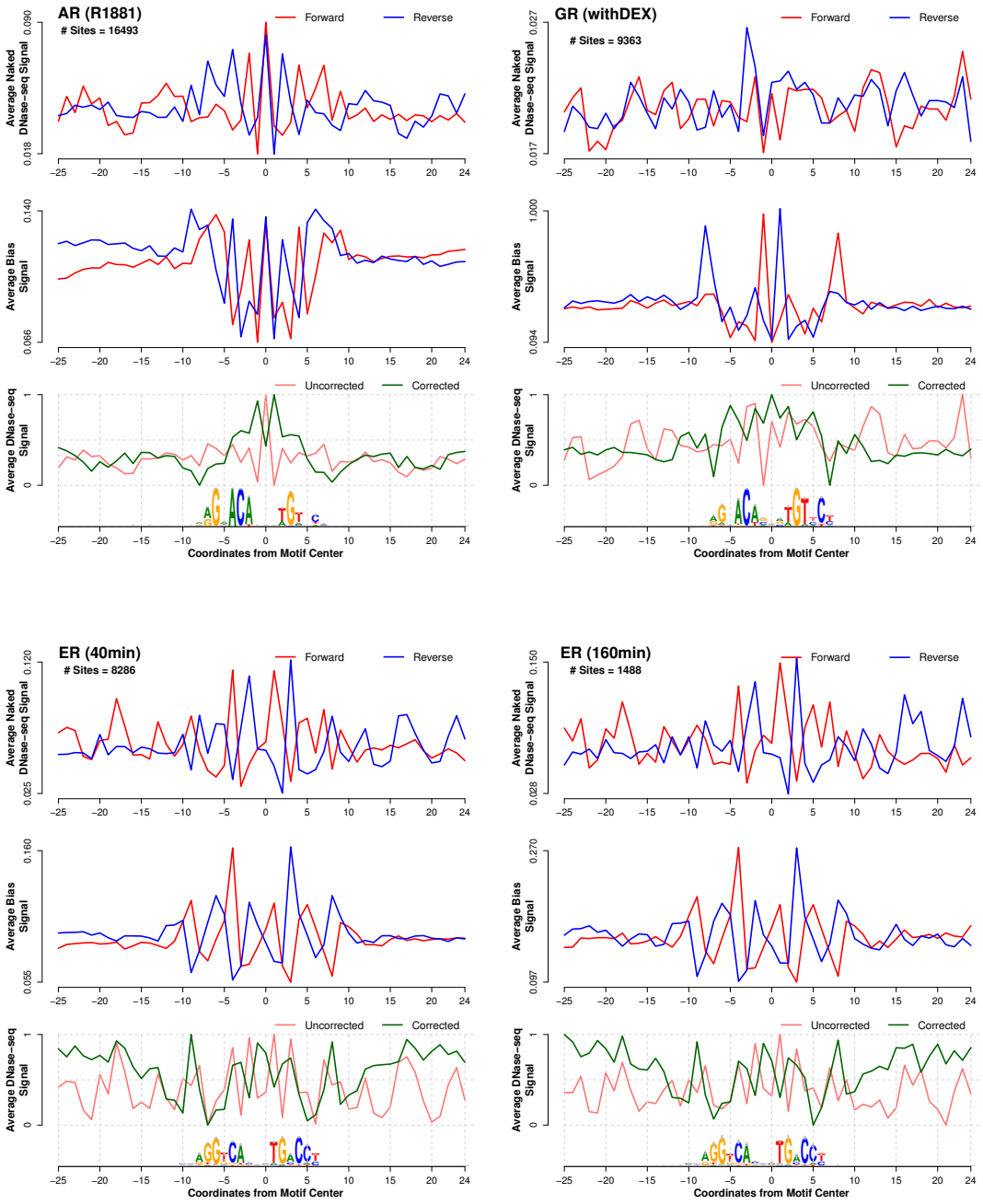

Figure 6: Average DNase-seq signals around nuclear receptor TFs with ChIP-seq evidence in LNCaP(DU), m3134(UW) and MCF-7(DU) cell types. In the top panel, we show the strand-specific average DNase-seq signal on deproteinized DNA experiments (MCF-7(DU) for data sets from single hit and IMR90(UW) for data 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.

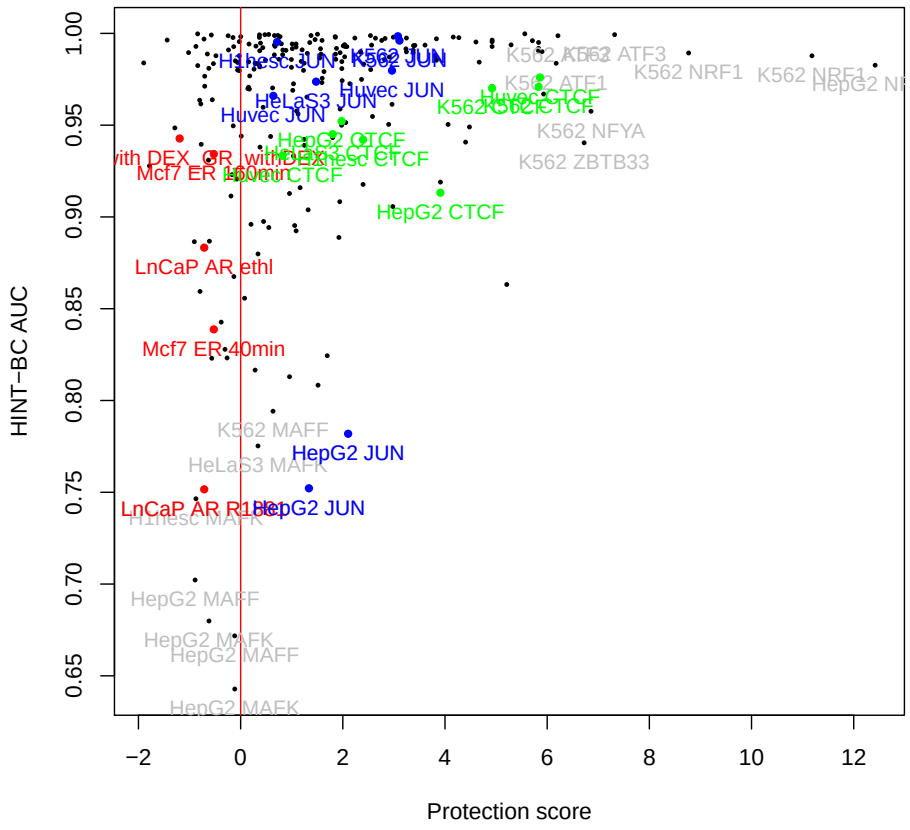

Figure 7: Scatter plot with the protection score (x-axis) vs. AUC of HINT-BC (y-axis) for 209 experiments. We indicate in red experiments with nuclear receptors AR, ER and GR (short residence time); in blue experiments with C-fos (medium residence time); in green experiments with CTFC (long residence time) and in gray experiments with either high protection score (> 6) or low AUC values (< 0.8).

.

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.

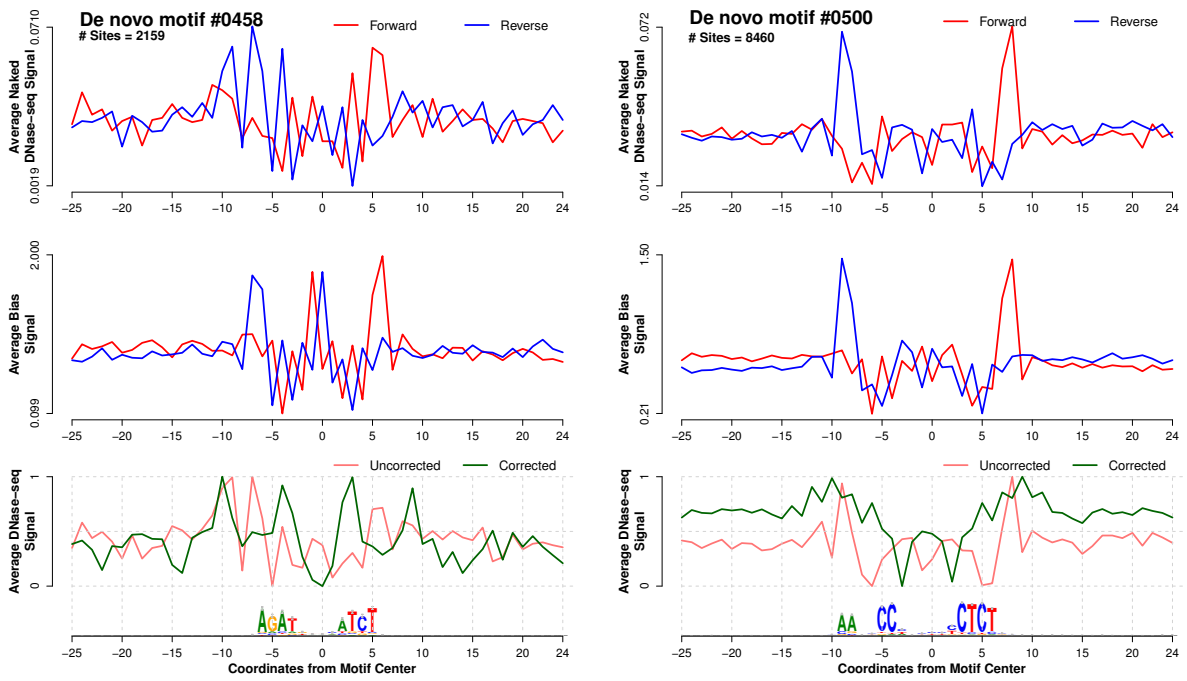

Figure 8: Average DNase-seq signals around binding sites of denovo motifs 0458 and 0500 on cell type H7-hESC. In the top panel, we show the strand-specific average DNase-seq signal on deproteinized DNA experiments (MCF-7 cell type); the middle panel shows the strand-specific estimated cleavage bias signal; 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.

.

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.

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

Natarajan, A., *et al.* (2012). Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Research*, **22**(9), 1711–1722.

Neph, S., *et al.* (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**(7414), 83–90.

Piper, J., *et al.* (2013). Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic acids research*, **41**(21), e201.

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.

Song, L. and Crawford, G. E. (2010). DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols*, **2010**(2), pdb.prot5384+.

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.

Wang, J., *et al.* (2012). Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Research*, **22**(9), 1798–1812.

Whitington, T., Perkins, A. C., and Bailey, T. L. (2009). High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Research*, **37**(1), 14–25.

Wilczynski, B., *et al.* (2009). Finding evolutionarily conserved cis-regulatory modules with a universal set of motifs. *BMC bioinformatics*, **10**(1), 82+.

Yardmc, 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+.