Analysis of computational footprinting methods for DNase sequencing experiments

Eduardo G. Gusmao[1,2], Manuel Allhoff[1,3], Martin Zenke[1,2], Ivan G. Costa[1,2,3,*].

[1] IZKF Computational Biology Research Group, RWTH Aachen University Medical School, Aachen, Germany.
[2] Department of Cell Biology, Institute of Biomedical Engineering, RWTH Aachen University Medical School, Aachen, Germany.
[3] Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Aachen, Germany.
* e-mail: ivan.costa@rwth-aachen.de

Final Version: Gusmao et al, Nature Methods, 13, 303–309, doi:10.1038/nmeth.3772
http://www.nature.com/nmeth/journal/v13/n4/pdf/nmeth.3772.pdf

Abstract: DNase-seq allows a nucleotide-level identification of transcription factor binding sites based on the computational search of footprint-like DNase I cleavage patterns on the DNA. Frequently, in high-throughput methods, experimental artifacts like DNase I cleavage bias impact the computational analysis of DNase-seq experiments. Here we performed a comprehensive and systematic study on the performance of computational footprinting methods. We evaluated 10 footprinting methods on a panel of DNase-seq experiments for their ability to recover cell-specific transcription factor binding sites. We show that three methods: HINT, DNase2TF and PIQ consistently outperform other evaluated methods. We demonstrate that correcting the DNase-seq signal for experimental artifacts significantly improves accuracy of computational footprints. We also propose a score to detect footprints arising from transcription factors with potentially short residence time.

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

NGS-based data are significantly affected by artifacts, which are inherent to the experimental protocols used[15,16,17]. An example is the DNase I sequence cleavage bias, which is due to DNase I having different binding affinities towards specific DNA sequences. He et al.[15] showed that sequence cleavage bias around TFBSs strongly affects the performance of a computational footprinting method[4,15] (footprint score; FS) in a TF-specific manner. They also indicated several TFs, such as nuclear receptors and *de novo* motifs found via computational footprinting[4], where the DNase-seq profile resembles their sequence cleavage bias estimate. Furthermore, they indicated that ranking putative TFBS by the number of DNase-seq reads around putative TFBSs[10,15] (tag count; TC) outperforms the ranking by FS. Another experimental aspect affecting the computational analysis of DNase-seq is the residence time of TF binding. Sung et al.[7] showed that short-lived TFs display a lower DNase I cleavage protection pattern, i.e. low number of DNase-seq reads surrounding the footprint. Moreover, they also noticed that nuclear receptors have DNase-seq profiles resembling their DNase I sequence cleavage bias estimates. While both studies[7,15] show the challenges imposed by cleavage bias and residence time, there have been a few attempts[7,12,15] to address these computationally.

There is no well-defined gold standard for the evaluation of footprinting methods. All work so far has used ChIP-seq of TFs in conjunction with motif-based predictions as ground truth. In short, motif-predicted binding

sites (MPBSs) supported by ChIP-seq peaks are positive examples (true TFBSs), while MPBSs without ChIP-seq support are negative examples (false TFBSs)[10]. This evaluation requires TF ChIP-seq experiments to be carried out on the very same cells as the DNase-seq experiment and has a few caveats. First, TF ChIP-seq peaks are also observed in indirect binding events[4,7,12,18]. Second, they have a lower spatial resolution than DNase-seq. Therefore, false TFBSs might be regarded as true TFBSs by proximity to a real TFBS of a distinct TF[15,17]. Recently, Yardımcı et al.[12] indicated that footprint quality scores, as measured by the footprint likelihood ratio (FLR), were significantly higher in cells where the TF was expressed. This observation indicates that comparing changes in expression and quality of footprints in a pairs of cells could provide an alternative footprint evaluation measure. Finally, with the exception of a few studies[8,11,12,13], comparative analyses evaluating footprinting methods were based on ChIP-seq of few (<12) TFs and with the exception of Gusmao et al.[8], a maximum of four competing methods were evaluated. Despite the importance of method evaluation[19], there is a clear lack of benchmark data, evaluation standards and studies performing a comprehensive analysis of computational footprinting methods.

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

RESULTS

Computational genomic footprinting methods

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

There are several other relevant characteristics for computational footprinting methods. A few methods allow the inclusion of additional genomic and/or experimental evidence like conservation scores[9], distance to transcription start sites[9] and histone modifications[8-10]. Only PIQ[11] supports the analysis of several DNase-seq data sets, i.e. experiments with replicates or time series. Another important feature is the correction of DNase-seq experimental artifacts, which is only supported by DNase2TF[7], HINT[8] variants (HINT-BC and HINT-BCN) and FLR[9]. While HINT-BC, HINT-BCN and DNase2TF use experimental bias statistics to pre-process DNase-seq profiles; FLR builds a "cleavage bias" model within their mixture model in a TF-specific manner. Most methods use base pair DNase-seq resolution as primary input[4-9,11-13]. One exception is Cuellar[10], which is based on smoothed DNase-seq signals of windows with 150 bps. Smoothing of base pair resolution profiles is performed by PIQ via the use of GP models[11]. BinDNase uses a greedy backward feature selection approach, which merges read counts of neighboring genomic positions[13]. Footprinting

methods also provide statistics to rank footprint predictions. Wellington[6] and DNase2TF[7] use read count statistics to provide *p*-values for each footprint. Several site-centric approaches provide either probabilities (BinDNase[13], Centipede[9] and PIQ[11]) or log-odds scores (FLR[12]) of footprints. Other methods use statistics such as FS (Neph[4]), PWM (Cuelar[10]) scores or TC (HINT[8]), to rank predicted footprints.

The availability, usability and scalability of software tools implementing the methods are also important features. Neph[4], HINT[8], PIQ[11] and Wellington[6] provide tutorials and software to run experiments with few command line calls. Of those, only HINT[8], PIQ[11] and Wellington[6] natively support standard genomic formats as input. Site centric methods Cuellar[10], BinDNase[13], Centipede[9] and FLR[12] require a single execution and input data per TF and cell, while segmentation methods require an execution per cell only. These site centric methods have computational demands 5 times (FLR and Cuellar) to 50 times (BinDNase and Centipede) higher than the slowest segmentation method (Wellington) on our analysis (**Supplementary Table 1**). The main method features are summarized in **Table 1** and described in the **Online Methods**.

Association of TF expression with footprint quality

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

Impact of experimental artifacts

To understand the nature of artifacts on DNase-seq experiments, we analyzed the sequence bias estimates on all 61 ENCODE Tier 1 and 2 DNase-seq data sets (**Supplementary Table 2**). These experiments include two main DNase-seq protocols, which differ on the number of DNase I digestion events necessary to generate DNA fragment (single-hit[2] and double-hit[3]). The sequence bias estimates can be defined as the ratio between the numbers of observed and expected DNase-seq reads starting at the middle of a particular DNA sequence of length k (k-mer)[15]. We use here two approaches. The "DHS sequence bias" considers the sequence bias estimates within DNase hypersensitive sites (DHSs) of each DNase-seq experiment. This approach captures DNase I cleavage, read fragmentation and sequence complexity bias of DHSs of each DNase-seq experiment[7,15]. The "naked DNA sequence bias" considers the sequence bias estimates within naked DNA DNase-seq experiments[12]. In this case, all DNA regions are open, therefore the sequence bias estimates will mainly capture the DNase I cleavage bias[12] (**Online Methods**). A clustering analysis of sequence bias estimates forms two clear groups, which splits experiments from single-hit and double-hit protocols (**Fig. 2**, **Supplementary Fig. 3**). This indicates that sequence biases are protocol-specific. Naked DNA sequence bias estimates forms a sub-cluster within estimates from the double-hit experiments. This highlights that DNase-seq experiments are influenced by more experimental artifacts than DNase sequence cleavage bias alone.

Next, we extended the analysis by He et al.[15] to evaluate the influence of sequence bias on all evaluated footprinting methods based on the AUC at 10% false positive rate (FPR). HINT was evaluated with DNase-seq signals corrected with either DHS sequence bias (HINT bias-corrected; HINT-BC) and naked DNA sequence bias (HINT bias-corrected on naked DNase-seq; HINT-BCN). Our analysis shows that only six out

of 14 evaluated methods (Wellington, Neph, Boyle, DNase2TF, Centipede and FS-Rank) present a significant negative Pearson correlation (r = –0.35, –0.32, –0.28, –0.28, –0.24 and –0.22, respectively) between their accuracy performance and amount of sequence bias (**Fig. 3a**; adjusted p-value < 0.05). Equivalent results are also observed on the same TFs and cellular conditions analyzed in He et al.[15] (**Supplementary Fig. 4**). Methods explicitly using *6*-mer sequence bias statistics (HINT-BC, HINT-BCN and FLR) or performing smoothing (Cuellar, BinDNase and PIQ) are not significantly influenced by sequence bias. Moreover, the performance of HINT-BC is the least affected by sequence bias (r = –0.06). Pairwise comparison of AUC at 10% FPR values of all three HINT variants (HINT-BC, HINT-BCN and HINT) indicates significant gain in all predictions with sequence bias correction (adjusted p-value < 10^{-30}; **Supplementary Fig. 5a**). There is no significant difference between HINT-BC and HINT-BCN, but we observe a higher AUC on HINT-BC on all but seven TFs. This indicates an advantage of DHS sequence bias correction for the footprint prediction problem.

As an example, we show sequence bias estimates, corrected and uncorrected DNase-seq average profiles around TFBSs with highest AUC gain between HINT-BC and HINT (**Fig. 3b and c**; **Supplementary Fig. 6**). The NRF1 and EGR1 DNase-seq profiles indicate that the bias-corrected signal fits better their sequence affinity than the uncorrected signal. We observe that *k*-mers with high DHS sequence bias have a high CG content (r > 0.8 in 11 out of 12 cell types; **Supplementary Fig. 7**). However, there is no significant correlation between CG content of MPBSs and either AUC values or differences of AUC from HINT-BC, HINT-BCN and HINT (p-value > 0.05; **Supplementary Fig. 5b**).

Comparative analysis of footprinting methods

Given its good performance[10,15], we evaluated the use of Tag Count (TC) as the ranking strategy instead of each method's own ranking for BinDNase, Centipede, Cuellar, DNase2TF, FLR, PIQ and Wellington. Previous to ranking by TC, site-centric methods required the definition of a minimum probability score to define active footprints. In all cases, using TC yielded higher AUC values (10% FPR) than using their intrinsic ranking metric (**Supplementary Fig. 8**). Concerning site-centric methods, the probability cutoff of 0.9 yielded highest AUCs, with exception of BinDNase (best at 0.8). These parameters will be used in the next evaluation analyses.

We next evaluated all the competing methods by measuring the AUC at 1%, 10% and 100% FPRs using the TF ChIP-seq data. AUC at lower FPRs favors methods with higher sensitivity in expense of specificity. We also estimated the AUPR, which is indicated for cases with imbalance of positive and negative examples[20], and the FLR-Exp metric. Interestingly, all TF ChIP-seq based metrics indicate a very similar ranking (r > 0.98; **Fig. 4a**). There is also a high agreement between FLR-Exp and other metrics (r > 0.88). HINT-BC has the highest FLR-Exp, AUC and AUPR values and significantly outperforms all methods with the exception of HINT-BCN (adjusted p-value < 0.01; **Supplementary Fig. 9**; **Supplementary Tables 3-6**). Ignoring HINT variants, the next top performing method is DNase2TF, which significantly outperforms all other methods with the exception of PIQ (adjusted p-value < 0.01). PIQ outperforms all of its lower ranked competitors but Wellington with AUC (1% FPR) and AUPR (adjusted p-value < 0.01). Concerning the performance of TC-Rank, we observe that the AUC values for 10% and 100% FPR are very close to other footprinting methods (**Fig. 4b, Supplementary Fig. 9**). This is not the case for AUC at 1% FPR or AUPR values. With the latter statistics, all methods but Centipede and Cuellar have significant superior performance than TC (p-value < 0.01; **Supplementary Tables 3-6**).

Transcription factor residence time

Despite the high average prediction values of top performing footprint methods, they consistently perform worst in a similar set of TFs, i.e. HINT-BC, DNase2TF and PIQ have 89% of TFs in common in the lower quartile of AUC at 10% FPR (**Supplementary Dataset 1**). This list includes nuclear receptors, which has low residence binding time[7] and display a lower DNase I cleavage protection pattern (**Supplementary Fig. 10**). To further investigate this, we propose a statistic inspired by the concepts presented in Sung et al.[7] to detect TFs with potential short residence time. The protection score measures the difference between the amounts

of DNase I digestion in the flanking regions and within the TFBS on bias-corrected DNase-seq signals. We use this statistic to analyze the predictive performance of methods on TFs with distinct residence time. For this, we used the comprehensive data set with 233 combinations of DNase-seq experiments and TFs (see **Online Methods**).

We observed that TFs with known short residence time on DNA, such as nuclear receptors AR[21], ER[22] and GR[23], present a negative protection score (**Fig. 5a**). TFs with intermediate and long residence time on DNA (C-JUN[24] and CTCF[25], respectively) present a positive protection score. The amount of protection is clearly reflected in the bias-corrected DNase-seq profiles (**Fig. 5b-d**). In addition, **Figure 5a** also reveals an association of the protection score and the AUC of HINT-BC. Overall, the protection score positively correlates with the AUC values of evaluated methods, such as TC (r = 0.19) and HINT-BC (r = 0.26), and negatively correlates (r = –0.49) with the sequence bias (adjusted p-value < 0.05). These results reinforce the concept that TFs with potential short residence time can be poorly detected via DNase-seq footprints.

DISCUSSION

Our comparative evaluation analysis indicates the superior performance (in decreasing order) of HINT, DNase2TF and PIQ in the prediction of active TFBSs in all evaluated scenarios. Moreover, tools implementing these methods were user friendly and had lower computational demands than other evaluated methods. Clearly, the choice of computational footprinting approaches should also be based on experimental design aspects. For example, PIQ is the only method supporting analysis of replicates and time-series. On the other hand, studies requiring footprint predictions for latter *de novo* motif analysis should use segmentation approaches as HINT or DNase2TF. In contrast to positive evaluations of the TC-Rank by previous works[10,15], we show that it has poor sensitivity performance as indicated by the AUC at low FPR levels. On the other hand, the TC statistic provides the best strategy to rank footprint predictions from other methods.

The refined DNase-seq protocol and experimental artifacts presented in He et al.[13] and TF binding time presented in Sung et al.[7] underscore that robust *in silico* techniques are required to correct for experimental artifacts and to derive valid biological predictions. The correction of DNase-seq signal with DHS sequence bias estimates virtually removes the effects of sequence bias artifacts on computational footprinting. We demonstrated that such correction can be performed prior to the execution of the computational footprinting method. On the other hand, ignoring experimental artifacts might lead to false predictions, as observed previously for predicted *de novo* motifs (**Supplementary Fig. 11**). Moreover, the simple protection score can indicate footprints of TFs with potential short binding time. Thus, footprint predictions of TFs with low protection score should be interpreted with caution.

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

Acknowledgements: This work was supported by the Interdisciplinary Center for Clinical Research (IZKF Aachen), RWTH Aachen University Medical School, Aachen, Germany (to E.G.G., M.A. and I.G.C.) and the Excellence Initiative of the German Federal and State Governments and the German Research Foundation through Grant GSC 111 (M.A. and I.G.C.).

Author Contributions: E.G.G., M.Z. and I.G.C. designed the research. E.G.G. wrote HINT program code. E.G.G., M.A. and I.G.C. analyzed data. E.G.G., M.Z. and I.G.C. wrote the manuscript.

Competing Financial Interests: The authors declare no competing financial interests.

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

Figure 1 | FLR-Exp evaluation metric. (**a**) FLR score distribution of footprints predicted with HINT-BC overlapping with MPBSs of selected TFs. These TFs have increasing expression in K562 (red) compared with H1-hESC cells (blue). The signed Kolmogorov-Smirnov (KS) statistic quantifies the separation of both distributions. The box plot depicts the distribution median value (middle dot) and first and third quartiles (box extremities). The whiskers represent the 1.5 IQR and external dots represent outliers (data greater than or smaller than 1.5 IQR). (**b**) Scatter plot with signed KS statistic and expression fold change for 143 TFs. There is a clear association between TF expression and KS statistic (r = 0.97, adjusted p-value < 10^{-10}).

Figure 2 | Clustering of bias estimates. Ward's minimum variance clustering based on pairwise Spearman correlation coefficient (r) from bias estimates of all ENCODE's Tier 1 and naked DNA DNase-seq data. DNase-seq experiments were based on single-hit (red), double hit (blue) protocols or naked DNA (yellow).

Figure 3 | Effects of sequence biases on methods. (**a**) Association between the performance of footprinting methods (relative to TC-Rank performance) and their sequence bias estimated for 88 TFs binding on cell types H1-hESC and K562. The *x*-axis represents the correlation between the uncorrected and bias signal (observed versus bias signal; OBS). The OBS is evaluated for each TF by measuring the uncorrected DNase-seq signal and the bias signal for every MPBS that overlaps a footprint from the evaluated method. Then, the Spearman correlation is evaluated between the average uncorrected and bias signals. Higher OBS values indicate higher bias. The *y*-axis represents the ratio between the AUC at 10% FPR for each evaluated method and the TC-Rank method; higher values indicate higher accuracy. (**b-c**) Average bias signal (top) and uncorrected/bias-corrected DNase-seq signal (bottom) for the TFs: (**b**) NRF1 and (**c**) EGR1. Signals in the bottom graph were standardized to be in the interval [0,1]. The motif logo represents all underlying DNA sequences centered on the TFBSs.

Figure 4 | Evaluation of computational footprinting methods. (**a**) Average rankings for the evaluated computational footprinting methods. The rankings are given for all evaluation criteria: FLR-Exp, TF ChIP-seq based AUC (at 100%, 10% and 1% FPR) and AUPR. (**b**) For all evaluated methods we show the FLR-Exp values (as a combination of all pairwise comparison within cell types H1-hESC, K562 and GM12878), median TF ChIP-seq based AUC (at 100%, 10% and 1% FPR) values and median AUPR values. HINT-BC, HINT-BCN, HINT, DNase2TF are ranked as top four methods by all evaluation metrics. All baseline methods (FS-Rank, PWM-Rank and TC-Rank) are in the bottom four positions of the ranks. Note that BinDNase could not be evaluated with the FLR-Exp, as it requires ChIP-seq data for training.

Figure 5 | Impact of transcription factor residence binding time on computational footprinting. (**a**) Scatter plot with the protection score (*x*-axis) versus TF ChIP-seq based AUC (at 10% FPR) of HINT-BC (*y*-axis) for 233 TFs binding on 11 cell types. We highlight nuclear receptors AR, ER and GR (short residence time, red); C-JUN (intermediate residence time, blue); CTCF (long residence time, green) and TFs with either high (> 6) protection score or low (< 0.8) AUC values (grey). (**b-d**) Average bias signal (top) and uncorrected/bias-corrected DNase-seq signal (bottom) for the TFs (**b**) ER, (**c**) C-JUN and (**d**) CTCF. Signals in the bottom graph were standardized to be in the interval [0,1]. The motif logo represents all underlying DNA sequences centered on the TFBSs.

Table 1 | Overview of methods. Main characteristics of the evaluated methods. Methods obtain a '+' sign for availability if they are public available. Boyle method is not public, but authors provide footprint predictions of a few cells. Concerning usability, methods natively supporting standard genomic files and being executed with few commands (≤ 3) display a '+' sign.

Name	Type	Algorithm	Bias Correction	Resolution/ Smoothing	Footprint Ranking	Availability	Usability	Others
BinDNase	SC	Logistic regression	None	Base pair / sliding window	Probability	+	-	Require TF ChIP-seq for training
Boyle	SEG	HMM	None	Base pair	None	-	-	
Centipede	SC	Bayesian mixture model	None	Base pair	Probability	+	-	Integrates histone and sequence data
Cuellar	SC	Weighted motif match	None	Sliding window	PWM score	+	-	
DNase2TF	SEG	Sliding window	*4*-mer (DHS sequence bias)	Base pair	*p*-values	+	+	
FLR	SC	Mixture model	*6*-mer (naked DNA sequence bias)	Base pair	Log-odds	+	-	Bias correction for each TF
HINT	SEG	HMM	*6*-mer (DHS sequence bias)	Base pair	TC	+	+	Integrates histones

Neph	SEG	Sliding window	None	Base pair	FS	+	-	
PIQ	SEG	GP / expectation propagation	None	Base pair / GP	Probability	+	+	Support replicates, time series
Wellington	SEG	Sliding window	None	Base pair	*p*-value	+	+	

METHODS

Data. DNase-seq aligned reads were obtained from ENCODE[1]. 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 naked DNA (deproteinized) DNase-seq experiments from cell types MCF-7 and K562 (DU)[12] and IMR90 (UW)[26]. 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[1]. 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)[1] track with exception of the following experiments, in which the enriched regions were obtained using bowtie-2[27] and MACS[28]. AR (R1881 treatment) ChIP-seq raw sequences for LNCaP cell type was obtained in gene expression omnibus (GEO) with accession number GSM353644[29]. 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[30]. GR (dexamethasone treatment) ChIP-seq raw sequences for m3134 cell type was obtained in the sequence read archive (SRA) under study number SRP004871[31]. 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)[1].

TF motifs (position frequency matrices; PFMs) were obtained from the Jaspar[32], Uniprobe[33] and Transfac[34] 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[4]. The accession codes for all TF ChIP-seq experiments and PFM IDs are available in the **Supplementary Datasets 1a and 2b-d**.

Sequence bias correction. *DNase I hypersensitivity 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

$$x = \langle x_{1,} \dots, x_N \rangle,$$

where N equals the number of bases in the genome and each $x_i \in N^0$ is the number of DNase-seq reads in which the 5′ position 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[35] was used to create smoothed DNase-seq signals using Parzen density estimates. Then, the smoothed signal x^{fseq} was fit to a gamma distribution,

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

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[1,35]. We refer to DHSs as a set of genomic intervals

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

where $h_i = [m, n]$ for $m < n \in N$ and L is the total number of DHSs. We ignore for simplicity of notation the fact that intervals are defined on distinct chromosomes or contigs.

Estimation of DNase-seq sequence bias. We use two approaches to estimate sequence bias of DNase-seq experiments: (1) aligned reads inside DHSs from DNase-seq experiments (termed "DHS sequence bias") following He et al.[15] and (2) all aligned reads for naked DNA experiments (termed "naked DNA sequence bias") following Yardımcı et al.[12]. The observed cleavage score for a *k*-mer w corresponds to the number of DNase I cleavage sites centered at w. The background cleavage score is defined by the total number of times w occurs. Then, the bias estimation is computed as the ratio between the observed and background cleavage scores. Mathematical formalizations of the bias estimation will be made based on the DHS sequence bias approach.

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

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

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

Similarly, the background cleavage score r_w can be evaluated as

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

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

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

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

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

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

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

The bias score b_i^s represents how many times the *k*-mer sequence $G^s\left[i - \frac{k}{2}..i + \frac{k}{2} + 1\right]$ was cleaved by the DNase I enzyme in comparison to its total occurrence in: (1) DHSs (DHS sequence bias approach); (2) the entire genome (naked DNA sequence bias approach). As observed by He et al.[15] 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.

DNase-seq sequence 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 ($\hat{b}_i^s$)[15]. These smoothed signals were based on a 50 bp window and can be written as

$$\hat{x}_i^s = \sum_{j=i-25}^{i+24} x_j^s$$

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

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

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

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

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

The corrected DNase-seq signal generated by **equation (1)** 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.

Computational footprinting methods. 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**.

Neph method. Neph et al.[4] used a simplified version of the segmentation method originally proposed in Hesselberth et al.[36]. 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. The footprint 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[4]. As predictions were not available for other DNase-seq experiments, we obtained the scripts and parameterization through Neph et al.[4] footprinting method code repository at https://github.com/StamLab/footprinting2012. 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[4]. Finally, we consider only predictions that occurred within DNase-seq hotspots, evaluated using the method first described in Sabo et al.[37]. We obtained all hotspots generated by Stamatoyannopoulous' lab in ENCODE[1] 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”.

Boyle method. Boyle et al.[5] designed a segmentation 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[38]. 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.[5] available at http://fureylab.web.unc.edu/datasets/footprints/. We will refer to this method as “Boyle”.

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'[9]. 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)[39] and Ensembl gene annotation from ENCODE[1,40] 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.[8].

Cuellar Method. Cuellar-Partida et al.[10] 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.[10]. 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[41] to obtain the predictions. We will refer to this method as “Cuellar”.

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.

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)[11]. 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 *common.r*. Briefly, MPBSs were generated with the script *pwmmatch.exact.r*. The DNase-seq signal was created using the script *bam2rdata.r*. And the footprints were detected with the script *pertf.r*.

Footprint mixture (FLR). Yardımcı et al.[12] 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 naked DNA sequence bias frequencies or estimated *de novo*. After successful estimation, MPBSs are scored with the log odds ratio for the footprint *versus* background model. The model takes DNase-seq cuts within a small window around the candidate profiles (± 25 bp) as input. DNase-seq sequence 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 naked DNA sequence 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”.

DNase2TF. DNase2TF is a segmentation approach based on a binomial *z*-score, which evaluates the depletion of DNase-seq reads around the candidate footprints[7]. At a second step, DNase2TF interactively merges close candidate footprints whenever they improve depletion scores. DNase2TF corrects for DNase I sequence 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 sequence bias correction. Other parameters were set to their default values: *minw = 6*, *maxw = 30*, *z_threshold = -2* and *FDR = 10^{-3}*.

HINT, HINT-BC and HINT-BCN. Recently, Gusmao et al.[8] have proposed the segmentation method HINT (HMM-based identification of transcription factor footprints) as an extension of Boyle method[5]. 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[32] and Uniprobe[33]. 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.[8].

We have performed two modifications to the method described in Gusmao et al.[8]. 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 “DHS sequence bias”, and HINT bias-corrected on naked DNA (HINT-BCN) for the “naked DNA sequence bias” estimation. These modifications required retraining of the HMM models. For this, we used the same manual annotation described in Gusmao et al.[8]. The novel methods and trained models are available as a command-line tool at www.costalab.org/hint-bc.

BinDNase. BinDNase is a site-centric method based on logistic regression to predict active/inactive MBBSs[13]. 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. 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.

Footprint score rank (FS-Rank). He et al.[15] 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.[4]. The FS statistic is defined as

$$FS_{MPBS_i} = -\left(\frac{n_{C,i}+1}{n_{R,i}+1} + \frac{n_{C,i}+1}{n_{L,i}+1}\right),$$

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

$$\begin{aligned} n_{C,i} &= \sum_{j=m_i}^{n_i} x_j \\ n_{R,i} &= \sum_{j=n_i}^{2n_i-m_i} x_j \\ n_{L,i} &= \sum_{j=2m_i-n_i}^{m_i} x_j \end{aligned} \quad (2)$$

Tag count rank (TC-Rank). 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.[15]. This can be written as

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

Both TC and FS can be used as quality scores for footprints. However, as a method (termed TC-Rank and FS-Rank) it consists on attributing these quality scores to each MPBS and evaluating the performance at these ranked MPBS. This observation also holds for the PWM-Rank method described below.

Evaluation. *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[42]. This was performed using Biopython[43]. Then, we created MPBSs by matching all PWMs against the human (hg19) and mouse (mm9) genomes using the fast performance motif matching tool MOODS[44]. 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.[45] 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 baseline method and will be referenced as "PWM-Rank".

Method comparison. Methods were evaluated using a site-centric approach[10], 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, 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 100%, 10% and 1% false positive rates (FPRs). 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)[20,46].

Segmentation 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 by 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. 8**).

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.[15], 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.[8] and Boyle et al.[5], which includes evaluation of PWM, 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.

Expression-based evaluation (FLR-Exp). As shown in Yardımcı et al.[12], 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 overlapping particular motifs and the expression of the TF.

We used limma[47] to perform between-array normalization on expression of H1-hESC, K562 and GM12878 cells and obtain fold change estimates. Then, we retrieved all non-redundant 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 using these PFMs.

Afterwards, we evaluated the FLR[12] score, TC[15] and FS[15] for the footprints of each 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.[12]. 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 comparison 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. 1 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**.

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.[7], 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:

$$PROT_{MPBS} = \sum_{i=1}^{N} \frac{(N_{R,i} - n_{C,i}) + (n_{L,i} - n_{C,i})}{2N}$$

where $MBPS = \{MPBS_1, \ldots, MPBS_N\}$ is set of binding sites for a given motif, $MPBS_i = [m_i, n_i]$ is the genomic location of the i-th binding site and $n_{C,i}$, $n_{L,i}$ and $n_{R,i}$ are the number of DNase-seq reads in the binding site, upstream and downstream flanking positions, respectively (see **equation (2)** for details).

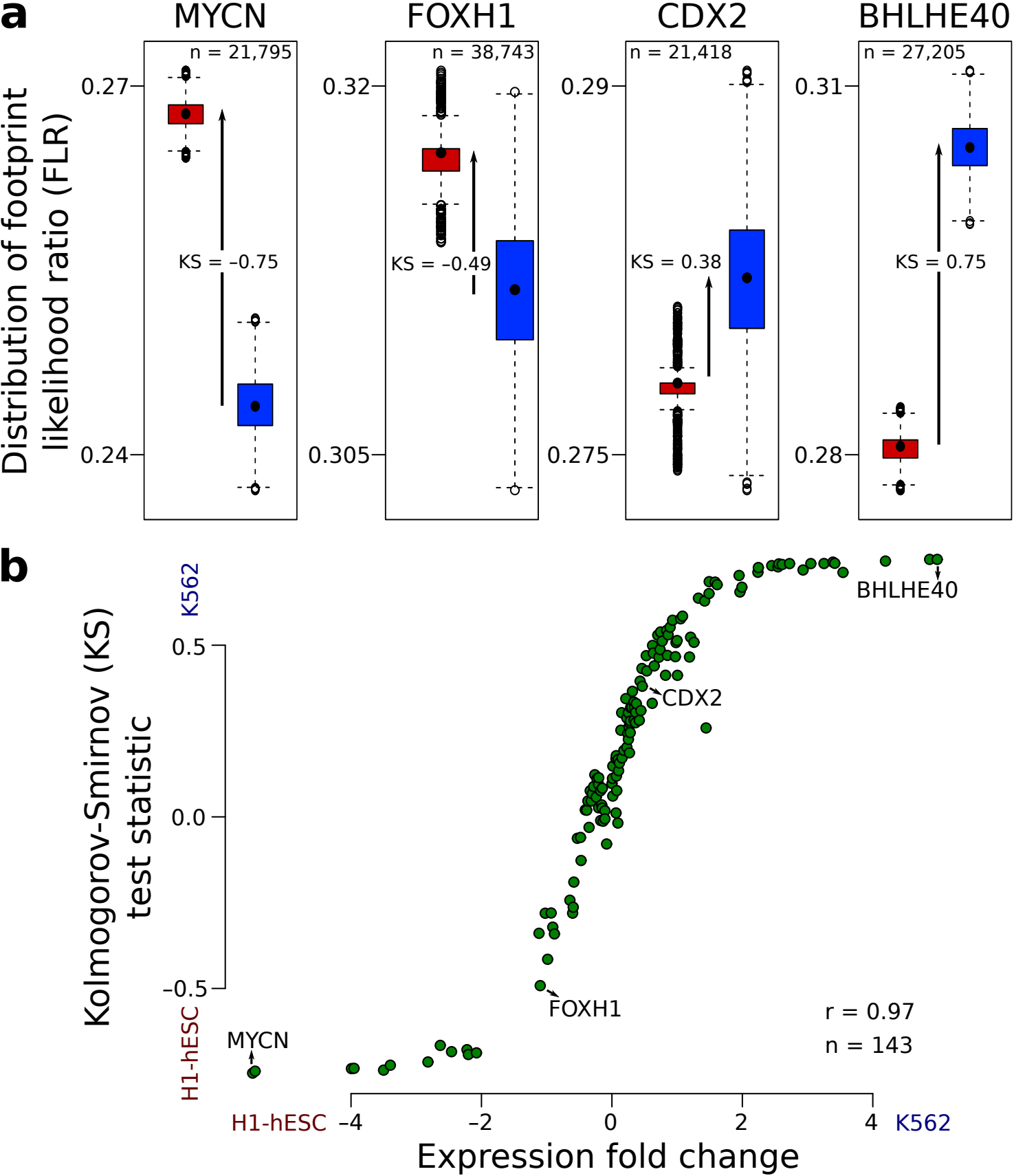
In short, the protection score indicates the average difference of DNase-seq counts in the flanking region and the DNase-seq counts within the MPBS. Positive values will indicate protection in the flanking regions, while values close to zero or negative indicate no protection. The protection score is a similar statistic as the FS[15]. The main difference is that the FS score measures the ratio between reads in flanking versus 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 DHS sequence bias and coverage differences. Results for protection scores are provided in **Supplementary Dataset 1**.

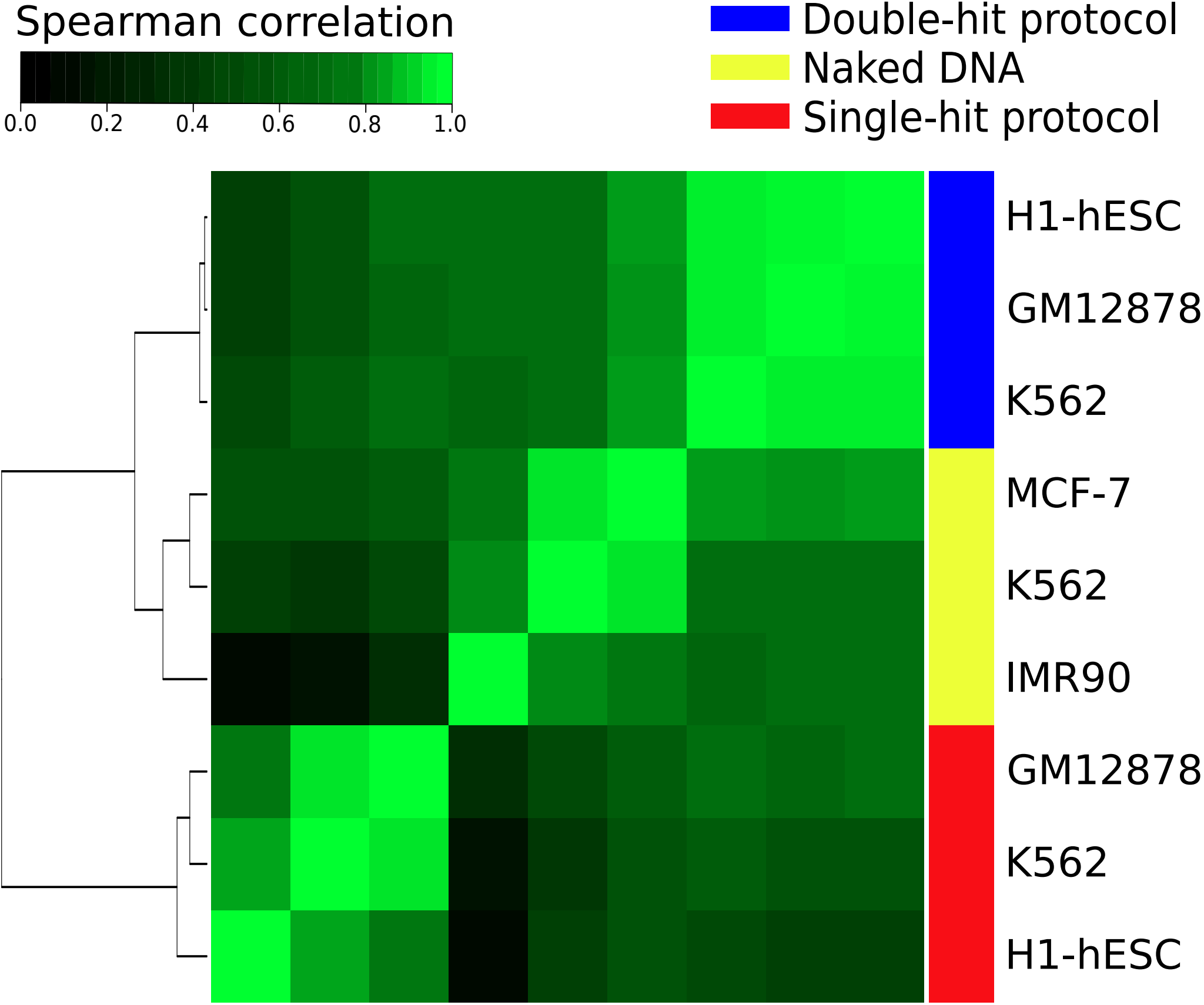
Statistical methods. The non-parametric Friedman-Nemenyi hypothesis test[48] was used to compare the AUC and AUPR of the methods regarding all data set combinations (TFs versus cell types). Such test provides a rank of the methods as well as the statistical significance of whether a particular method was outperformed. All correlations are based on Spearman values. All reported *p*-values have been corrected with the Benjamini and Hochberg method[49].

Code Availability. Software, custom code, benchmarking data, DNase-seq sequence bias estimates and further graphical results are available at www.costalab.org/hint-bc. The HINT, HINT-BC and HINT-BCN softwares can be directly accessed through the regulatory genomics toolbox website at www.regulatory-genomics.org/hint/.

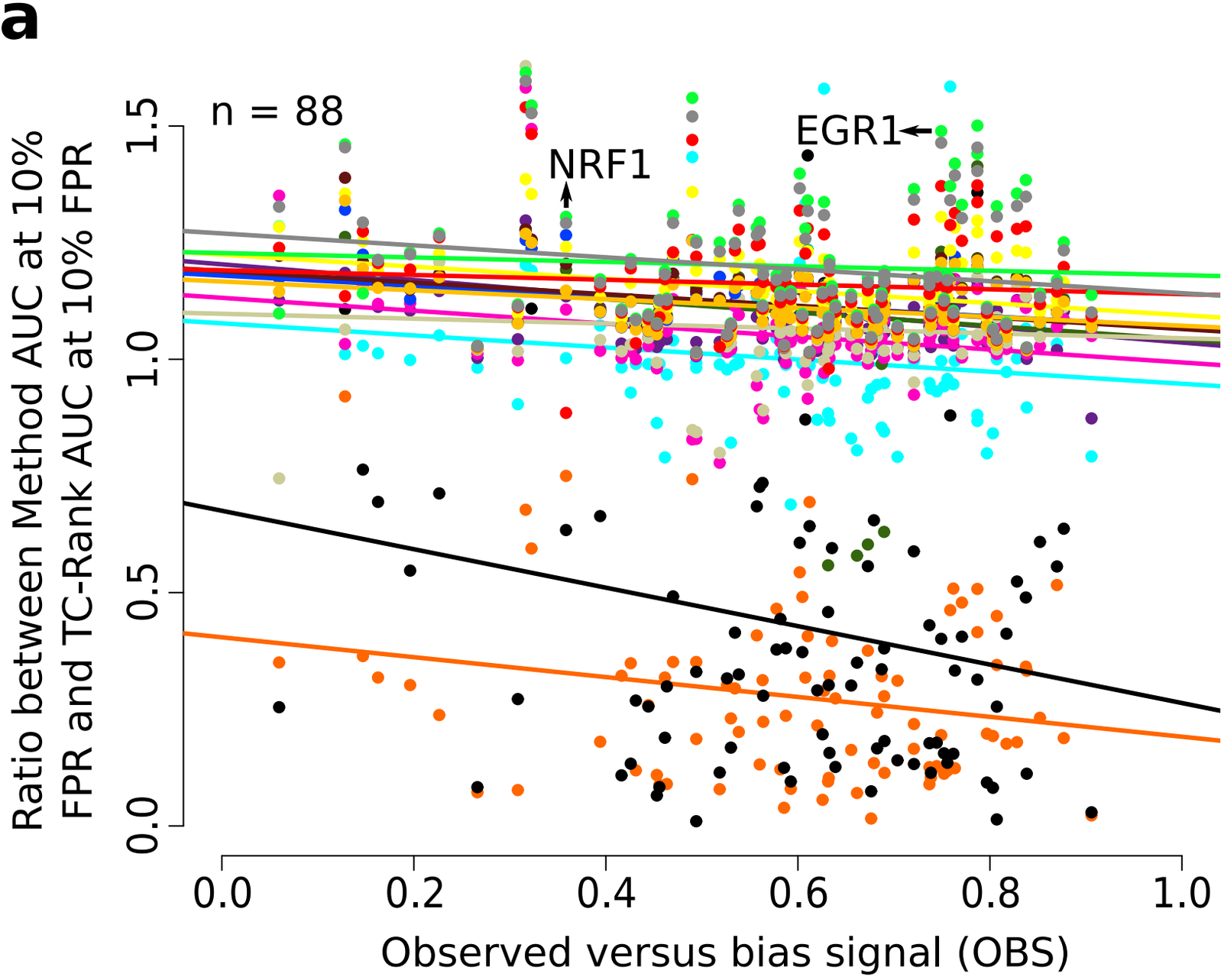
[26] Lazarovici, A. et al. Probing DNA shape and methylation state on a genomic scale with DNase I. *Proc. Natl. Acad. Sci. USA* **110**, 6376–6381 (2013).
[27] Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
[28] Zhang, Y. et al. Model-based analysis of ChIP-seq (MACS). *Genome Biol.* **9**, R137+ (2008).
[29] Yu, J. et al. An integrated network of androgen receptor, polycomb and TMPRSS2-ERG gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443–454 (2010).
[30] Guertin, M.J, Zhang, X., Coonrod, S.A. & Hager, G.L. Transient estrogen receptor binding and p300 redistribution support a squelching mechanism for estradiol-repressed genes. *Mol. Endocrinol.* **28**, 1522–1533 (2014).
[31] John, S. et al. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).
[32] Mathelier, A. et al. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42**, D142–D147 (2014).
[33] Robasky, K. & Bulyk, M. L. UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res.* **39**, D124–D128 (2011).
[34] Matys, V. et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
[35] Boyle, A.P. Guinney, J., Crawford, G.E. & Fury, T.S. F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**, 2537-2538 (2008).
[36] Hesselberth, J. R. et al. Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nat. Methods* **6**, 283–289 (2009).
[37] Sabo, P. J. et al. Discovery of functional noncoding elements by digital analysis of chromatin structure. *Proc. Natl. Acad. Sci. USA* **101**, 16837–16842 (2004).
[38] Madden, H.H. Comments on the Savitzky-Golay convolution method for least-squares fit smoothing and differentiation of digital data. *Anal. Chem.* **50**, 1383–1386 (1978).
[39] Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
[40] Hubbard, T. et al. The ensembl genome database project. *Nucleic Acids Res.* **30**, 38–41 (2002).
[41] Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
[42] Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16–23 (2000).
[43] Cock, P.J.A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
[44] Korhonen, J., Martinmäki, P., Pizzi, C. Rastas, P. & Ukkonen, E. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**, 3181–3182 (2009).
[45] Wilczynski, B., Dojer, N., Patelak, M. & Tiuryn, J. Finding evolutionarily conserved cis-regulatory modules with a universal set of motifs. *BMC Bioinform.* **10**, 82+ (2009).
[46] Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **27**, 861–874 (2006).
[47] Ritchie, M. E. et al. *limma* powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
[48] Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **7**, 1–30 (2006).

[49] Benjamini, Y. & Hochberg, Y. et al. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).


a
MYCN
FOXH1
CDX2
BHLHE40
n = 21,795
n = 38,743
n = 21,418
n = 27,205
0.27
0.24
0.32
0.305
0.29
0.275
0.31
0.28
KS = –0.75
KS = –0.49
KS = 0.38
KS = 0.75
Distribution of footprint likelihood ratio (FLR)
b
K562
H1-hESC
0.5
0.0
–0.5
Kolmogorov-Smirnov (KS) test statistic
BHLHE40
CDX2
FOXH1
MYCN
r = 0.97
n = 143
H1-hESC
–4
–2
0
2
4
K562
Expression fold change

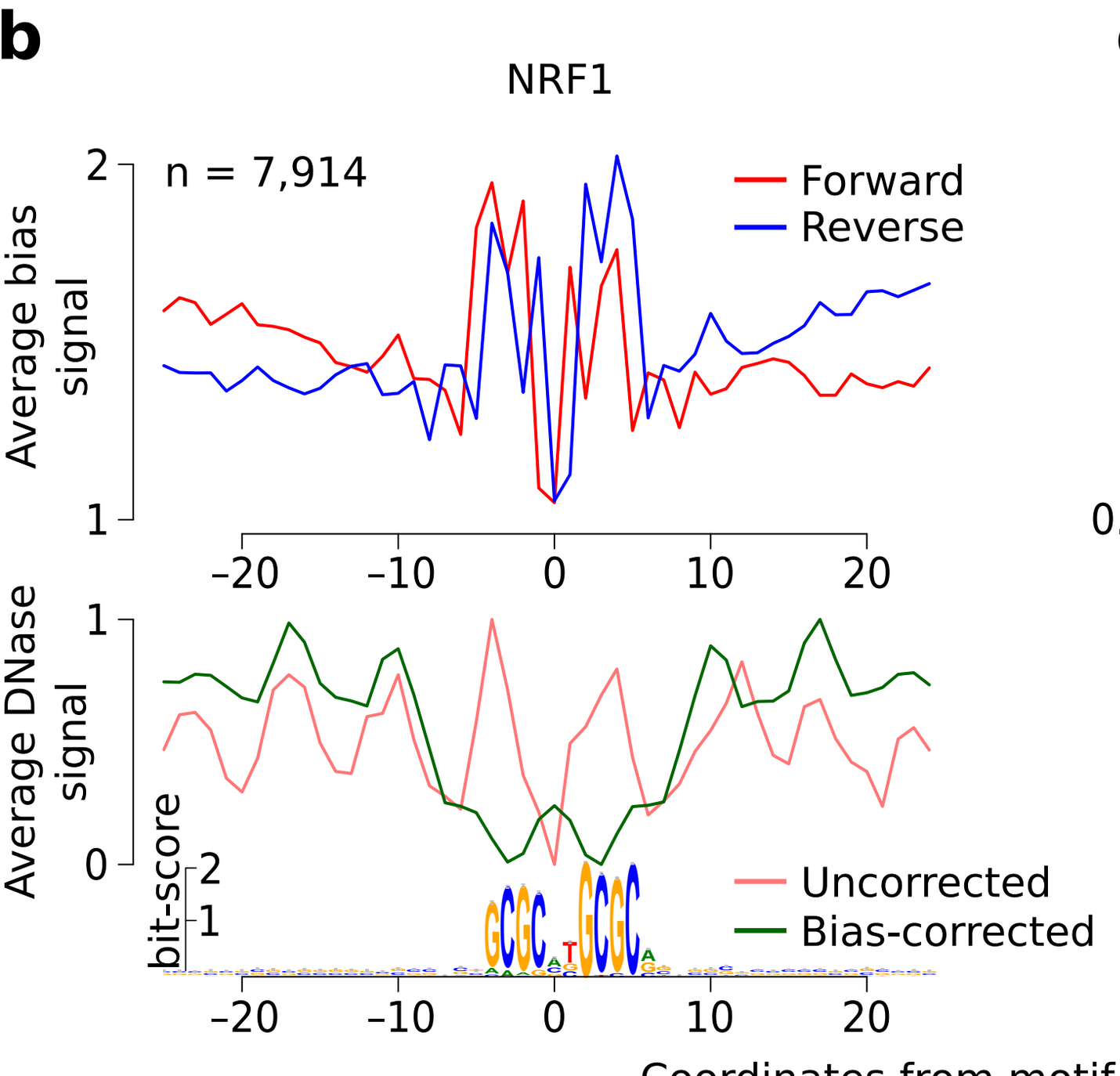

Spearman correlation
0.0
0.2
0.4
0.6
0.8
1.0
Double-hit protocol
Naked DNA
Single-hit protocol
H1-hESC
GM12878
K562
MCF-7
K562
IMR90
GM12878
K562
H1-hESC

a

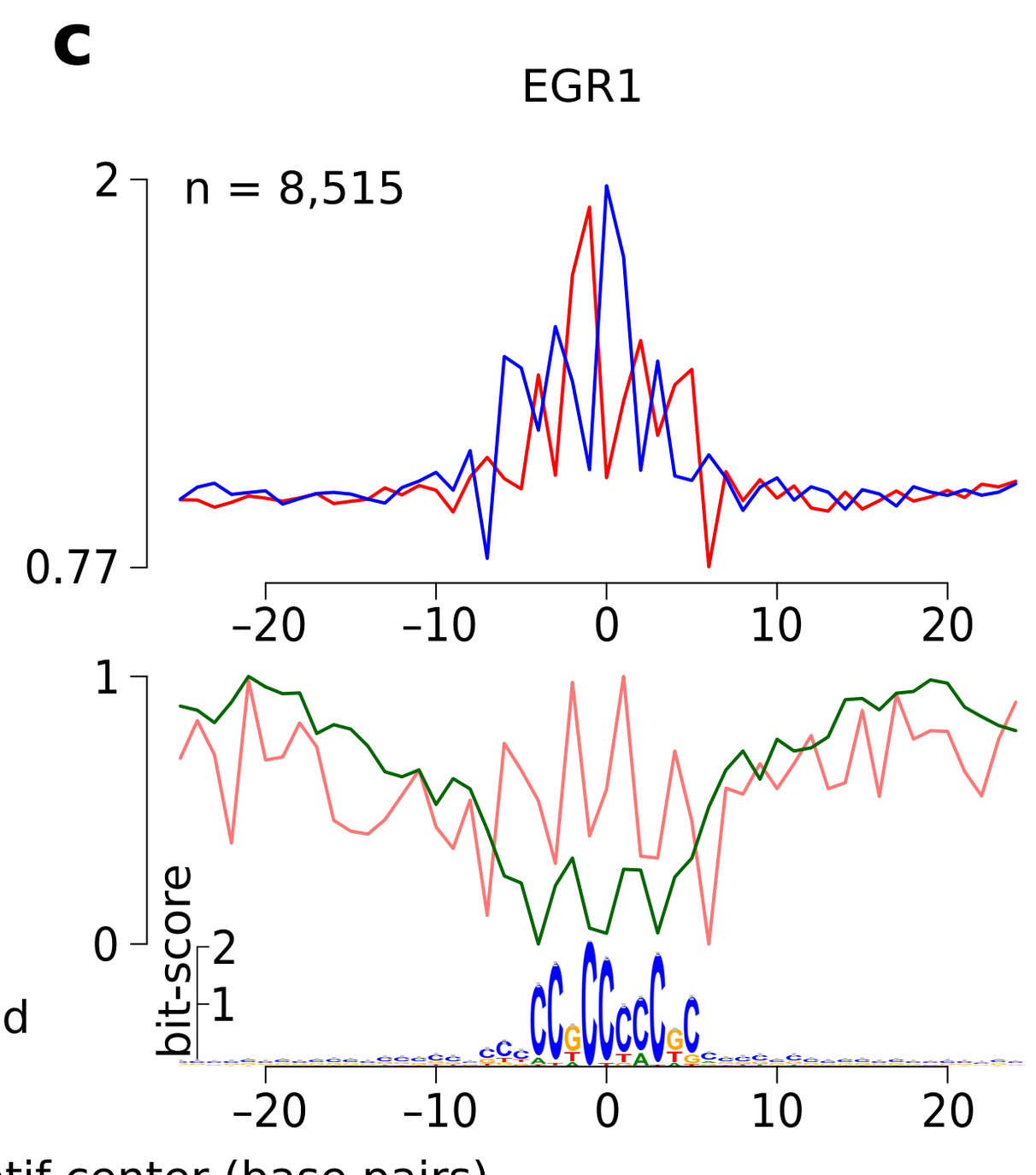


	Correlation	p-value
Wellington	–0.35	0.0009
Neph	–0.32	0.0025
Boyle	–0.28	0.0072
DNase2TF	–0.28	0.0085
Centipede	–0.24	0.0246
FS-Rank	–0.22	0.0394
PWM-Rank	–0.2	0.0561
HINT-BCN	–0.18	0.0834
PIQ	–0.18	0.0857
Cuellar	–0.16	0.1308
FLR	–0.13	0.2162
BinDNase	–0.09	0.4252
HINT	–0.08	0.438
HINT-BC	–0.06	0.5473

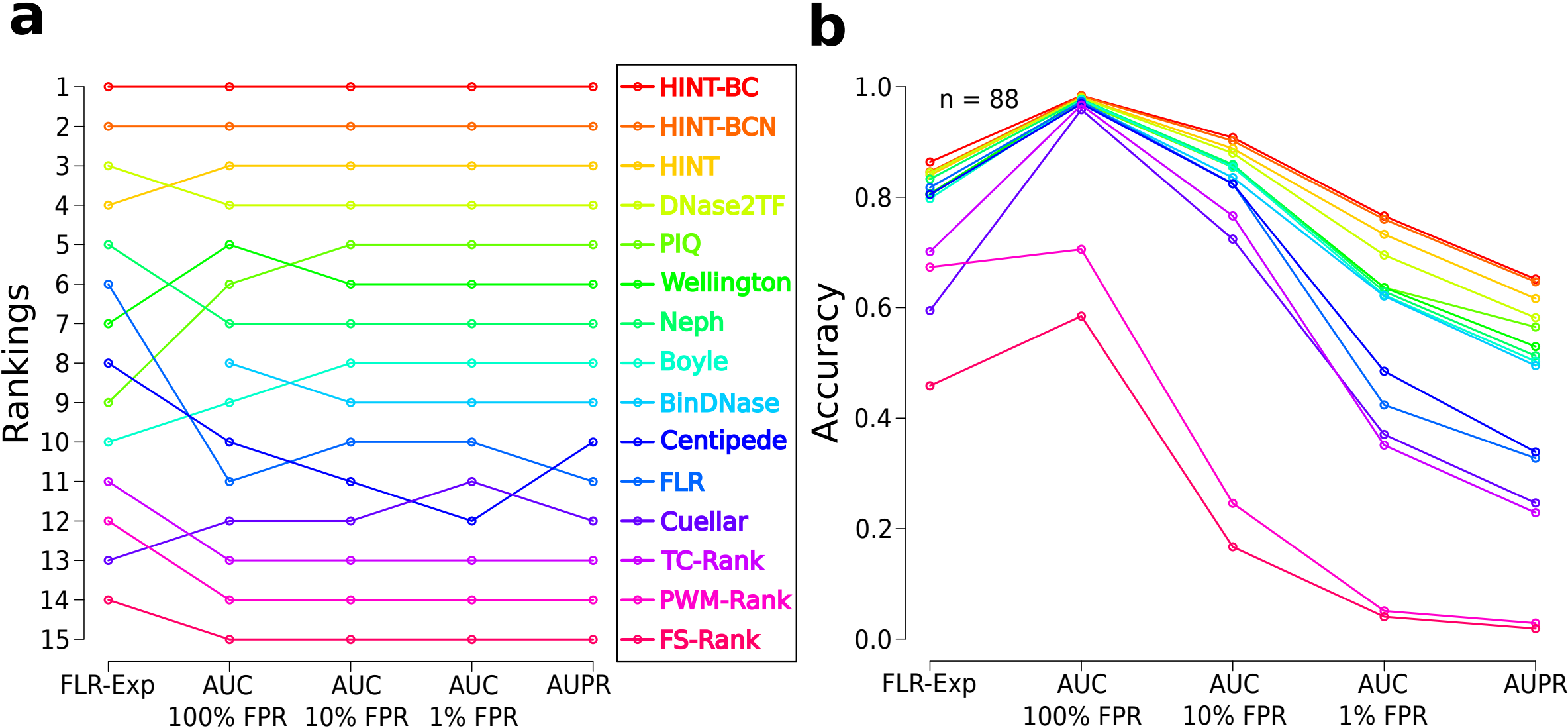
b



c



Coordinates from motif center (base pairs)


a
b
Rankings
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
FLR-Exp
AUC 100% FPR
AUC 10% FPR
AUC 1% FPR
AUPR
HINT-BC
HINT-BCN
HINT
DNase2TF
PIQ
Wellington
Neph
Boyle
BinDNase
Centipede
FLR
Cuellar
TC-Rank
PWM-Rank
FS-Rank
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
n = 88
FLR-Exp
AUC 100% FPR
AUC 10% FPR
AUC 1% FPR
AUPR

a

Short-lived
Intermed.-lived
Long-lived

HINT-BC AUC (percentage)

n = 233

C-JUN(K562)
C-JUN(H1-hESC)
C-JUN(K562)
C-JUN(Huvec)
C-JUN(HeLa-S3)
C-JUN(Huvec)
GR(m2134)
ER-160m(MCF-7)
AR-ethl(LNCaP)
ER-40m(MCF-7)
MAFF(K562)
C-JUN(HepG2)
MAFK(HeLa-S3)
AR-R1881(LNCaP)
C-JUN(HepG2)
MAFK(H1-hESC)
MAFF(HepG2)
MAFK(HepG2)
MAFF(HepG2)
MAFK(HepG2)
ATF3(K562)
NRF1(K562)
NRF1(K562)
NRF1(Hep
CTCF(Huvec)
CTCF(K562)
CTCF(K562)
CTCF(HepG2)
CTCF(HeLa-S3)
CTCF(H1-hESC)
CTCF(Huvec)
CTCF(HepG2)
NF-YA(K562)
ZBTB33(K562)

Protection score

b

ER (40min)

n = 8,286

Average bias signal

Average DNase signal

bit-score

protection

Coordinates from motif center (base pairs)

c

C-JUN

n = 982

Forward
Reverse

Uncorrected
Bias-corrected

protection

d

CTCF

n = 46,540

protection

Coordinates from motif center (base pairs)