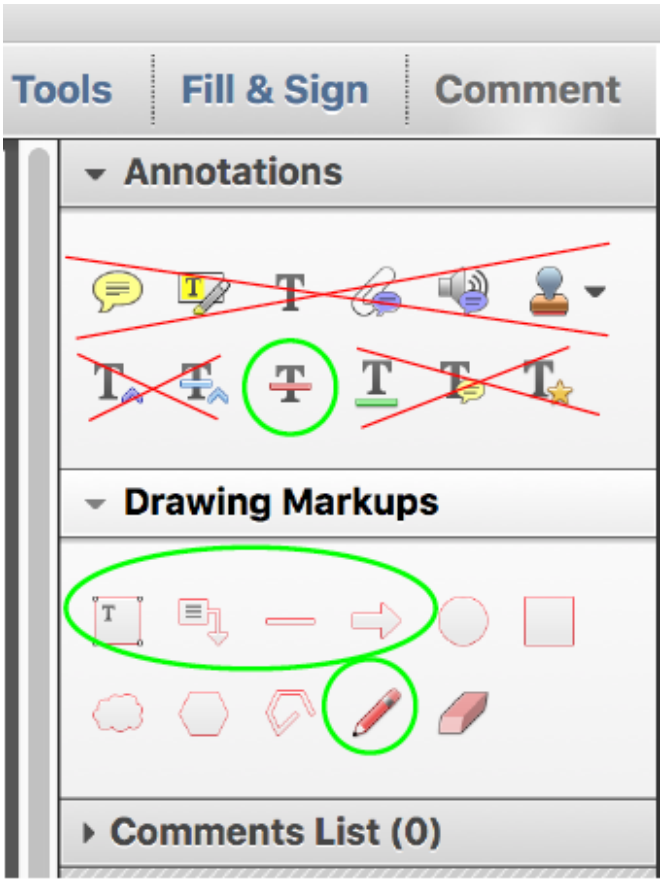
Instructions for reviewing your article proof
You will need to use Adobe Reader version 7 or above to make comments on this proof, available free from http://get.adobe.com/reader/.

Responding to the author query form: Please address each query on the query list, on the proof or in a separate query response list.



Editing the proof: Edit the manuscript as necessary, using only the circled tools in the Drawing or Annotation menu.

IMPORTANT: All edits must be visible in full on a printed page, or they will be lost in production. Do not use any tool that does not show your changes directly on a printed page.

Use Text boxes to add text changes and the drawing tools to indicate insert points or graphics changes. Extensive directions should be addressed in a text box. While most figure edits should be marked on the page, extensive visual changes to figures (e.g., adding scale bars or changes to data) should be accompanied by a new figure file with a written explanation of the changes.

Special characters can be inserted into text boxes by pasting from a word processing document or by copying from the list below.

α β χ δ ε φ γ η ι φ κ λ μ ν ο π θ ρ σ τ υ ϖ ω ξ ψ ζ

Α Β Σ Δ Ε Φ Γ Η Ι ϑ Κ Λ Μ Ν Ο Π Θ Ρ Σ Τ Υ ς Ω Ξ Ψ Ζ

Å Δ ≥ ≤ ≠ × ± 1° 3′ ↑ ↓ →←

Analysis of computational footprinting methods for DNase sequencing experiments

Q1 Eduardo G Gusmao[1,2], Manuel Allhoff[1,3], Martin Zenke[1,2] & Ivan G Costa[1–3]

DNase-seq allows nucleotide-level identification of transcription factor binding sites on the basis of a computational search of footprint-like DNase I cleavage patterns on the DNA. Frequently in high-throughput methods, experimental artifacts such as DNase I cleavage bias affect the computational analysis of DNase-seq experiments. Here we performed a comprehensive and systematic study on the performance of computational footprinting methods. We evaluated ten footprinting methods in 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 outperformed the other evaluated methods and that correcting the DNase-seq signal for experimental artifacts significantly improved the accuracy of computational footprints. We also propose a score that can be used to detect footprints arising from transcription factors with potentially short residence times.

Q1. Please remove affiliation number 1 from author "Martin Zenke". This author should have only the affiliation number 2.

Please change "improved" by "improves".

Next-generation sequencing combined with genome-wide mapping techniques such as DNase-seq has contributed greatly to our understanding of gene regulation and chromatin dynamics[1–3]. DNase-seq allows for nucleotide-level identification of transcription factor binding sites (TFBSs). This can be done via a computational search of footprint-like regions with low numbers of DNase I cuts surrounded by regions with high numbers of cuts[2,3]. A number of computational footprinting methods have been proposed in 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].

Next-generation sequencing–based data are significantly affected by artifacts, which are inherent to the experimental protocols used[15–17]. An example is the DNase I sequence cleavage bias, which is due to the different binding affinities of DNase I toward 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 f… s… t… a… comes than ranking by FS (number of DNase-seq reads inside and around a motif-predicted binding site (MPBS)). Another experimental aspect affecting the computational analysis of …me of TF binding. Sung *et al.*[7] showed … lower DNase I cleavage-protection …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. Although both of the aforementioned 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 challenges computationally.

There is no well-defined gold standard for the evaluation of footprinting methods. All work so far has used chromatin immunoprecipitation followed by sequencing (ChIP-seq) of TFs in conjunction with motif-based predictions as ground truth. In short, MPBSs supported by ChIP-seq peaks are positive examples (true TFBSs), whereas 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, the signal generated after TF ChIP-seq computational processing has a lower spatial resolution than DNase-seq signal. Therefore, false TFBSs might be regarded as true TFBSs if they are in close proximity to a real TFBS of the same TF[15,17]. Recently, Yardimci *et al.*[12] indicated that footprint quality scores, as measured by the footprint likelihood ratio (FLR), are significantly higher in cells where the TF in question is expressed. This observation indicates that comparing changes in the expression and quality of footprints in pairs of cells could provide an alternative footprint evaluation measure. Finally, with the exception of a few studies[8,11–13], comparative analyses evaluating footprinting methods have been

[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. Correspondence should be addressed to I.G.C. (ivan.costa@rwth-aachen.de).

RECEIVED 31 JULY 2015; ACCEPTED 27 JANUARY 2016; PUBLISHED ONLINE XX XX 2016; DOI:10.1038/NMETH.3772

Better to use: "PIQ<ref 11> (protein interaction quantification)" to follow the same format as HINT (in the line above). Please change <ref 11> by the superscript citation.

based on ChIP-seq of four (≤13) TFs and with the exception of [illegible] ChIP-seq–based approach, we evaluated ten computational footprinting methods—Neph[4], Boyle[5], Wellington[6], DNase2TF[7], HINT[8] (HMM-based identification of TF footprints), CENTIPEDE[9], Cuellar[10], protein interaction quantification[11] (PIQ), FLR[12] and BinDNase[13]—with respect to their accuracy in recovering TFBSs supported by 88 ChIP-seq TF experiments in two cell types (H1-hESC and K562) with area under the receiver operating characteristic curves (AUCs) and area under the precision-recall curves (AUPRs). 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 with 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 evaluated 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 as either segmentation[4–8] or site-centric[9–13] methods. Several segmentation methods use a 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 is based on hidden Markov models, 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 either bound or unbound. Most site-centric methods are based on unsupervised statistical methods such as mixture models (FLR[12]), Bayesian mixture models (CENTIPEDE[9]) and a combination of Gaussian process and expectation propagation (PIQ[11]). An alternative site-centric approach proposed by Cuellar *et al.*[10] uses DNase-seq profiles as the prior distribution for the detection of MPBSs. BinDNase is a supervised site-centric method based on logistic regression[13]. We also evaluated simple statistics as baseline methods, ranking MPBSs by position weight matrix (PWM-Rank) bit-score[10], by the ratio of the number of DNase-seq reads inside an MPBS to that around it (FS-Rank)[4,15] and by the number of DNase-seq reads around an 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 such as 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 supported only by DNase2TF[7], HINT[8] variants (HINT-BC and HINT-BCN) and FLR[9]. Whereas HINT-BC, HINT-BCN and DNase2TF use experimental bias statistics to pre-process DNase-seq profiles, FLR builds a 'cleavage bias' model within the 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 from 150-bp windows. Smoothing of base pair–resolution profiles is performed by PIQ via the use of Gaussian process models[11]. BinDNase uses a greedy backward feature-selection approach that merges read counts of neighboring genomic positions[13]. Footprinting methods also provide statistics that can be used 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]), position weight matrix scores (Cuellar[10]) and TC (HINT[8]) to rank predicted footprints.

The availability, usability and scalability of the software tools implementing the methods are also important features. Neph[4], HINT[8], PIQ[11] and Wellington[6] provide tutorials and software for running experiments with few command-line calls. Of those, only HINT[8], PIQ[11] and Wellington[6] natively support standard genomic formats as input. The site-centric methods Cuellar[10], BinDNase[13], CENTIPEDE[9] and FLR[12] require a single execution and input data source per TF and cell, whereas segmentation methods require an execution per cell only. These site-centric methods had computational demands 5 times (FLR and Cuellar) to 50 times (BinDNase and CENTIPEDE) higher than those of the slowest segmentation method (Wellington) in 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

Yardimci *et al.*[12] indicated that the FLRs of candidate footprints are significantly higher in cells where the relevant TF is being expressed. We expanded on this idea by evaluating whether differences in the 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 the top-performing methods in comparisons between pairs of H1-hESC, K562 and GM12878 cells (**Fig. 1b** and **Supplementary Fig. 1**). We also evaluated the use of the TC and FS metrics as quality scores instead of the FLR and found that they had lower average correlation values (TC, $r = 0.35$; FS, $r = 0.73$; **Supplementary Fig. 2**). We therefore opted to use the FLR as the quality measure for footprints for this evaluation procedure. We used the correlation between FLR score difference and expression fold change, which we refer to as FLR-Exp, to rank footprinting methods; higher FLR-Exp values indicate better performance. The FLR-Exp evaluation methodology requires only expression data and is therefore more generally applicable than TF ChIP-seq–based evaluation. However, unlike TF ChIP-seq evaluation, the FLR-Exp approach cannot evaluate footprint predictions for individual TFs.

Impact of experimental artifacts

To understand the nature of artifacts in DNase-seq experiments, we analyzed the sequence bias estimates of all 61 ENCODE tier 1 and 2 DNase-seq data sets (**Supplementary Table 2**). These

If possible, to reduce space in this table, please use abbreviations like: "DHS-BC" for "DHS sequence bias" and "NKD-BC" for "naked DNA sequence bias". Add the necessary description in the caption if this is done.

Table 1 | Overview of methods

Name	Type	Algorithm	Bias correction	Resolu / smo				
BinDNase	SC	Logistic regression	None	Base pa / wi				hIP-seq / ng
Boyle	SEG	HMM	None	Bas				
CENTIPEDE	SC	Bayesian mixture model	None	Bas				istone 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)	Bas				Integrates histones
Neph	SEG	Sliding window	None	Bas				
PIQ	SEG	GP, expectation propagation	None	Base				Supports replicates, time series
Wellington	SEG	Sliding window	None	Base pair	*P* value	+	+	

HINT has two variants. It may use "DHS sequence bias" or "naked DNA sequence bias". Please add the latter.

Main characteristics of the evaluated methods. Under "Availability," "+" denotes that the method is publically available. The Boyle method is not public, but the authors provide footprint predictions for a few cells. Under "Usability," "+" denotes that the method natively supports standard genomic files and can be executed with few commands (≤3). SC, site centric; SEG, segmentation approach; HMM, hidden Markov model; PWM, position weight matrix; GP, Gaussian process.

experiments included two main DNase-seq protocols that differ in the number of DNase I digestion events necessary to generate DNA fragments (single-hit[2] and double-hit[3]). Sequence bias estimates can be defined as the ratio of observed to expected DNase-seq reads starting at the middle of a particular DNA sequence of length k (k-mer)[15]. We used two approaches. The DNase hypersensitive site (DHS) sequence bias approach considers the sequence bias estimates within DHSs in each DNase-seq experiment. This approach captures the DNase I cleavage, read fragmentation and sequence-complexity bias of DHSs in each DNase-seq experiment[7,15]. The naked DNA sequence bias approach considers the sequence bias estimates in naked DNA DNase-seq experiments[12]. In this case, all DNA regions are open, and therefore the sequence bias estimates capture mainly the DNase I cleavage bias[12] (Online Methods). Our clustering analysis of [illegible] formed two clear groups containing ex[illegible]nd double-hit protocols, respectivel[illegible]y **Fig. 3**). This indicated that sequer[illegible]ic. Naked DNA sequence bias estim[illegible]thin estimates from the double-hit [illegible]Nase-seq experiments are influenced [illegible]cts than DNase sequence cleavage bias alone.

If consistent with journal format, please change the dot here for a colon punctuation.

Next we extended the analysis by He *et al.*[15] to evaluate the influence of sequence bias on all evaluated footprinting methods on the basis of the AUC at a 10% false positive rate (FPR). In this analysis, the Spearman correlation is calculated for the distribution of TF accuracy for each method in comparison to the distribution of TFs' observed (uncorrected) DNase-seq signal versus the bias signal. We evaluated the correlation between

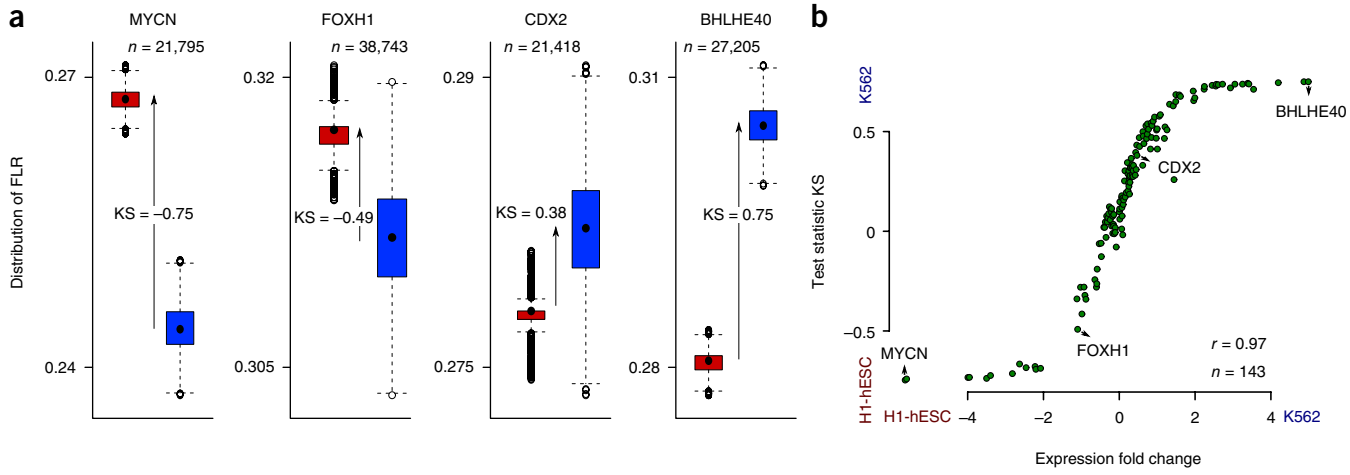


Figure 1 | FLR-Exp evaluation metric. (**a**) FLR score distribution of footprints predicted with HINT-BC overlapping with MPBSs of selected TFs. These TFs have increased expression in K562 cells (red) compared with H1-hESC cells (blue). The Kolmogorov-Smirnov (KS) statistic quantifies the separation of the two distributions. Box plots depict distribution median values (center dots) and first and third quartiles (box edges). The whiskers represent 1.5 times the interquartile range, and external dots represent outliers (data greater than or less than 1.5 times the interquartile range). (**b**) KS statistics and expression fold change for 143 TFs in H1-hESC and K562 cells. There is a clear association between TF expression and KS statistic ($r = 0.97$, adjusted $P < 10^{-10}$).

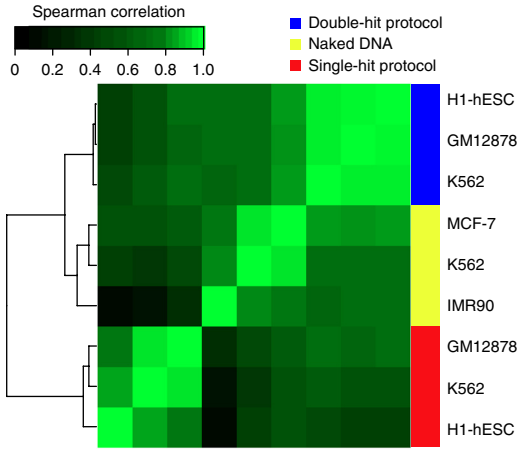


Figure 2 | Clustering of bias estimates. Ward's minimum variance clustering based on the pairwise Spearman correlation coefficient (r) from bias estimates of selected ENCODE tier 1 and naked DNA DNase-seq data. DNase-seq experiments were based on single-hit or double-hit protocols or on naked DNA.

the uncorrected and bias signals for each TF by measuring the uncorrected DNase-seq signal and the bias signal for every MPBS that overlapped a footprint from the evaluated method. Then we evaluated the Spearman correlation between the average uncorrected and bias signals. We evaluated HINT using DNase-seq signals corrected with either DHS sequence bias (HINT bias-corrected (HINT-BC)) or naked DNA sequence bias (HINT bias-corrected on naked DNase-seq (HINT-BCN)). Our analysis showed that only six out of ten evaluated methods (Wellington, Neph, Boyle, DNase2TF, CENTIPEDE and FS-Rank) presented a significant negative Spearman correlation ($r = -0.35, -0.32, -0.28, -0.28, -0.24$ and -0.22, respectively) between their accuracy and the amount of sequence bias (**Fig. 3a**; adjusted P value < 0.05). Equivalent results were also observed for the same TFs and cellular conditions analyzed by He *et al.*[15] (**Supplementary Fig. 4**). Methods explicitly using 6-mer sequence bias statistics (HINT-BC, HINT-BCN and FLR) or including smoothing (Cuellar, BinDNase and PIQ) were not significantly influenced by sequence bias. Moreover, the performance of HINT-BC was the least affected by sequence bias ($r = -0.06$). Pairwise comparison of AUCs at 10% FPR for all three HINT variants (HINT-BC, HINT-BCN and HINT) indicated significant gain in all predictions with sequence bias correction (adjusted $P < 10^{-30}$; **Supplementary Fig. 5a**). There was no significant difference between HINT-BC and HINT-BCN, but we observed a higher AUC with HINT-BC for 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 and corrected and uncorrected DNase-seq average profiles around TFBSs with the highest AUC gain between HINT-BC and HINT (**Fig. 3b**,**c** and **Supplementary Fig. 6**). The NRF1 and EGR1 DNase-seq profiles indicate that the bias-corrected signal fit their sequence affinity better than the uncorrected signal. We observed that k-mers with high DHS sequence bias had a high CG content ($r > 0.8$ in 11 out of 12 cell types; **Supplementary Fig. 7**). However, there was no significant correlation between the CG content of MPBSs and either AUC values or differences in AUC among HINT-BC, HINT-BCN and HINT ($P > 0.05$; **Supplementary Fig. 5b**).

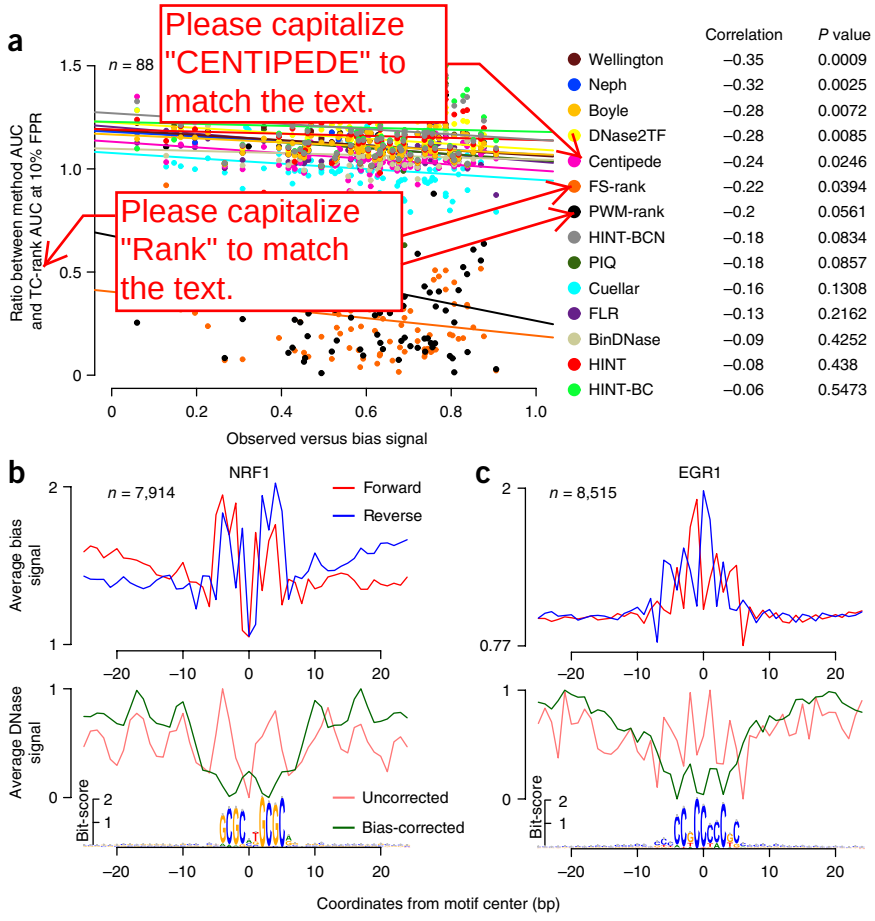


Figure 3 | Effects of sequence bias on methods. (**a**) Association between the performance of footprinting methods (relative to TC-Rank performance) and their sequence bias estimated for 88 TFs binding on H1-hESC and K562 cells. The x-axis represents the correlation between the uncorrected and bias signals (OBS). Higher OBS values indicate higher bias. The y-axis represents the ratio between the AUC at 10% FPR for each evaluated method and for the TC-Rank method; higher values indicate higher accuracy. The diagonal lines in the plot represent the linear regression's best fit. (**b**,**c**) Average bias signal (top) and uncorrected and bias-corrected DNase-seq signals (bottom) for the TFs (**b**) NRF1 and (**c**) EGR1. DNase signals were standardized to be in the interval [0,1]. The motif logos represent all underlying DNA sequences centered on the TFBSs.

Comparative analysis of footprinting methods

Given its good performance[10,15], we evaluated the use of the TC as the ranking strategy instead of each method's own ranking for BinDNase, CENTIPEDE, Cuellar,

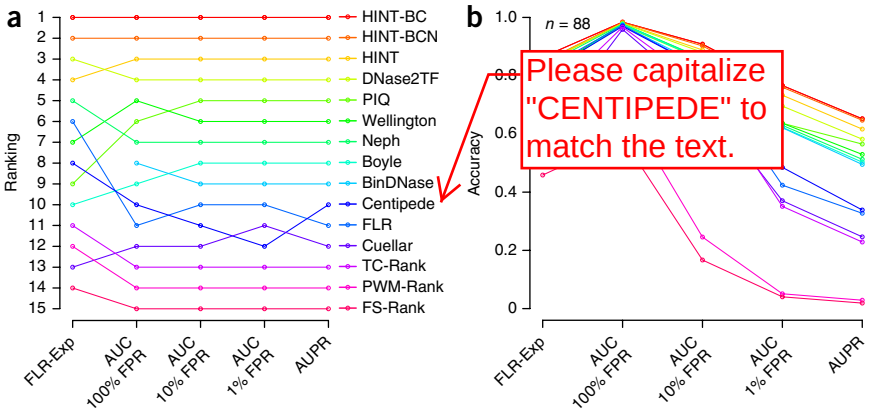


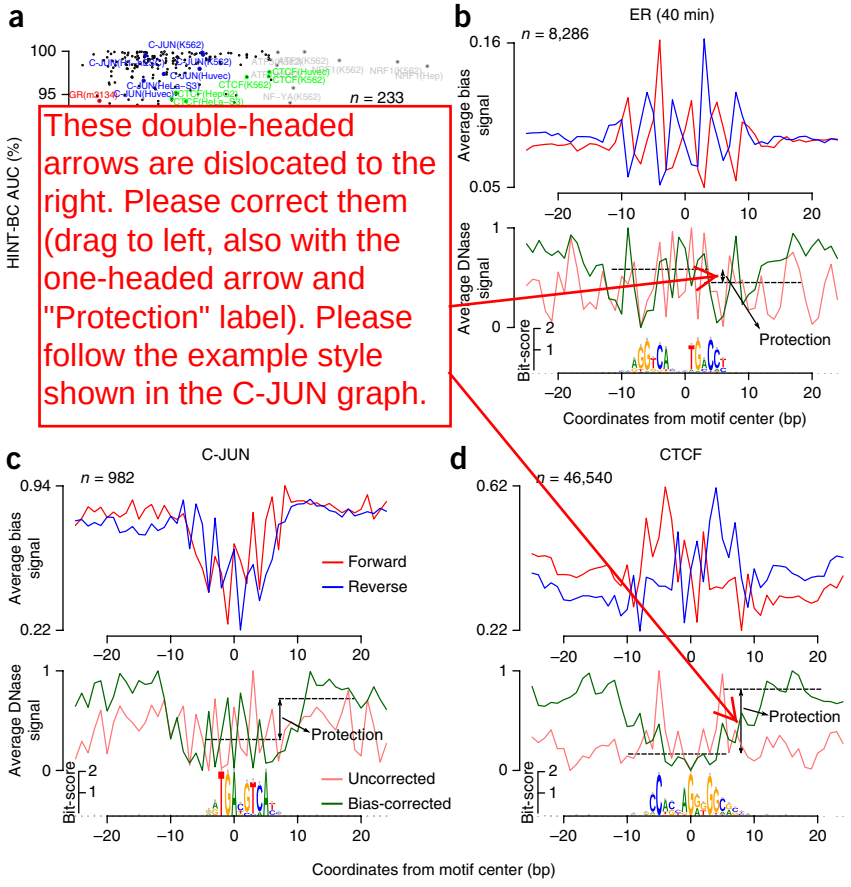
Figure 4 | Evaluation of computational footprinting methods. (**a**) Average rankings for the evaluated computational footprinting methods. The rankings are given for all evaluation criteria: FLR-Exp, TF ChIP-seq–based AUC (at 100%, 10% and 1% FPR) and AUPR. (**b**) FLR-Exp values (as a combination of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

DNase2TF, FLR, PIQ and Wellington. Prior 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 did using the intrinsic ranking metric (**Supplementary Fig. 8**). A probability cutoff of 0.9 yielded the highest AUCs for site-centric methods, with the exception of BinDNase (highest AUC at 0.8). These parameters were used in our subsequent evaluation analyses.

We next evaluated all the competing methods by measuring the AUC at 1%, 10% and 100% FPR using TF ChIP-seq data. AUCs at lower FPRs favored methods with higher sensitivity at the expense of specificity. We also estimated the AUPR, which is indicated for cases with an imbalance between positive and negative examples[20], and the FLR-Exp metric. All TF ChIP-seq–based metrics indicated a very similar ranking ($r > 0.98$; **Fig. 4a**). There was also strong agreement between FLR-Exp and other metrics ($r > 0.88$). HINT-BC had the highest FLR-Exp, AUC and AUPR values and significantly outperformed all other methods except HINT-BCN (adjusted $P < 0.01$; **Supplementary Fig. 9** and **Supplementary Tables 3–6**). When we ignored HINT variants, the next top-performing method was DNase2TF, which significantly outperformed all other methods but PIQ (adjusted $P < 0.01$). PIQ outperformed all of its lower-ranked competitors except for Wellington in terms of AUC (1% FPR) and AUPR (adjusted $P < 0.01$). Concerning the performance of TC-Rank, we observed that its AUC values for 10% and 100% FPR were very close to those for other footprinting methods (**Fig. 4b** and **Supplementary Fig. 9**). This was not the case for values of AUC at 1% FPR or of AUPR, for which all methods but CENTIPEDE and Cuellar had significantly superior performance relative to that of TC ($P < 0.01$; **Supplementary Tables 3–6**).

TF residence time

Despite the high average prediction values of the top-performing footprint methods, they consistently performed worst with a similar set of TFs; HINT-BC, DNase2TF



Q6. These labels are allowed to overlap since the position of the dots is the important information. Furthermore the TFs are being clearly explicited in the figure caption.

Average bias signal (top) and uncorrected and bias-corrected DNase-seq signals (bottom) for the TFs (**b**) ER (**c**) C-JUN and (**d**) CTCF. DNase signals were standardized to be in the interval [0,1]. Motif logos represent all underlying DNA sequences centered on the TFBSs.

and PIQ had 89% of TFs in common in the lower quartile of the AUC at 10% FPR (**Supplementary Data Set 1**). This list included nuclear receptors, which have low residence binding times[7] and display a lower DNase I cleavage-protection pattern (**Supplementary Fig. 10**). To further investigate this, we proposed a statistic inspired by the concepts presented by Sung *et al.*[7] to detect TFs with potential short residence times, the protection score, which measures the difference between the amounts of DNase I digestion in the flanking regions and in the TFBS on bias-corrected DNase-seq signals. We used this statistic to analyze the predictive performance of different methods with TFs with distinct residence times. For this we used a comprehensive data set with 233 combinations of DNase-seq experiments and TFs (Online Methods).

We observed that TFs with known short residence times on DNA, such as nuclear receptors AR[21], ER[22] and GR[23], presented a negative protection score (**Fig. 5a**). TFs with intermediate and long residence times on DNA (C-JUN[24] and CTCF[25], respectively) presented a positive protection score. The amount of protection was clearly reflected in the bias-corrected DNase-seq profiles (**Fig. 5b–d**). We also noted association of the protection score and the AUC in HINT-BC (**Fig. 5a**). Overall, the protection score positively correlated with the AUC values of evaluated methods such as TC ($r = 0.19$) and HINT-BC ($r = 0.26$) and negatively correlated ($r = -0.49$) with the sequence bias (adjusted $P < 0.05$). These results reinforce the concept that TFs with potentially short residence times are poorly detected via DNase-seq footprints.

DISCUSSION

Our comparative analysis indicates the superior performance of (in decreasing order) 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 approach should also be based on experimental design aspects. For example, PIQ is the only method supporting analysis of replicates and time series. For studies requiring footprint predictions for *de novo* motif analysis, one should use a segmentation approach such as HINT or DNase2TF. In contrast to positive evaluations of TC-Rank in previous works[10,15], we show that it has poor sensitivity as indicated by the AUC at low FPR levels. The TC statistic provided the best strategy for ranking footprint predictions from other methods.

The refined DNase-seq protocol and experimental artifacts presented by He *et al.*[15] and the TF binding time presented by Sung *et al.*[7] underscore the idea 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 have demonstrated that such correction can be performed before the execution of the computational footprinting method. It should still be noted, however, that 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 potentially short binding times. Thus, footprint predictions for TFs with low protection scores 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 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 the transparency and reproducibility of research on computational methods for DNase-seq data.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

Q2. Acknowledgements are correct.

This work was supported [illegible] (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 (grant GSC 111 to 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

Q3. Competing financial interests are correct.

The authors declare no competing f[illegible]

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. Yardimci, G.G., Frank, C.L., Crawford, G.E. & Ohler, U. 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.A. & Liu, X.S. 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. & van 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. The difficulty of a fair comparison. *Nat. Methods* **12**, 273 (2015).
20. Davis, J. & Goadrich, M. The relationship between precision-recall and ROC curves. *Proc. 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, Z.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).

EDITORIAL SUMMARY

AOP: This comparison of ten computational methods for detecting transcription factor binding sites in DNase hypersensitive regions in the genome determines which methods work consistently well, how DNase-seq experimental artifacts should be corrected for and which score is best for ranking methods.

ONLINE METHODS

Data. DNase-seq aligned reads were obtained from ENCODE[1]. To perform the computational footprint experiments, we obtained data for H1-hESC, HeLa-S3, HepG2, HUVEC, K562, LNCaP and MCF-7 cells from Crawford's lab at Duke University (labeled "DU" in the following) and for H7-hESC, HepG2, HUVEC, K562 and m3134 cells from Stamatoyannopoulous's lab at the University of Washington (labeled "UW" in the following). We also used naked DNA (deproteinized) DNase-seq experiments with MCF-7 and K562 cells (DU)[12] and with IMR90 cells (UW)[26]. Experiments on DU cells followed the single-hit protocol, and experiments on UW cells followed the double-hit protocol. In addition, to perform the DNase-seq bias estimation clustering, we used all tier 1 and tier 2 cell types from ENCODE[1]. **Supplementary Table 2** presents a full description of DNase-seq data.

TF ChIP-seq enriched regions (peaks and summits) were obtained from the ENCODE analysis working group[1] track with the exception of the following experiments, in which the enriched regions were obtained using Bowtie 2 (ref. 27) and MACS[28]. AR (R1881 treatment) ChIP-seq raw sequences for m3134 cells were obtained from the Sequence Read Archive (SRA) under study number SRP004871 (ref. 31). All organism-specific data (DNase-seq and ChIP-seq) are based on human genome build 37 (hg19), except the DNase-seq data for m3134 and ChIP-seq data for GR, which were based on mouse genome build 37 (mm9). Chromosome Y was removed from all analyses. Expression profiles of H1-hESC, K562 and GM12878 cells were obtained from ENCODE[1] (GSE12760 and GSE14863).

ensembl/encode/supplementary/integration_data_jan2011/byDataType/footprints/jan2011/de.novo.pwm. The accession codes for all TF ChIP-seq experiments and PFM IDs are available in **Supplementary Data Sets 1a** and **2b–d**.

where N is 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 maps to position i. We also generated strand-specific counts X^s, where $s \in \{+,-\}$ describes the strand the read was mapped to.

DHSs were estimated on the basis of the DNase I raw signal. First we used the F-seq software[35] to create smoothed DNase-seq signals using Parzen density estimates. Then we fit the smoothed signal x^{fseq} to a gamma distribution,

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

by evaluating κ and θ on the basis of mean and s.d. estimates. Finally, we found the enriched regions (DHSs) by establishing a cutoff based on a P value of 0.01 (refs. 1,35). We refer to DHSs as a set of genomic intervals

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

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

Estimation of DNase-seq sequence bias. We used two approaches to estimate sequence bias in 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 Yard1mc1 *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 are made on the basis of the DHS sequence bias approach.

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

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

where $1(.)$ is an indicator function.

Similarly, the background cleavage score r_w can be evaluated as

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

Finally, the cleavage bias b_i^s for a genomic position $k + 1 \le i \le N - k + 1$, given that $w = G^s[i - (k/2)\ldots i + (k/2)]$, can be calculated as

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

where O^s indicates the number of ... in DHSs,

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

Please change all occurrences of this name to "Yardimci" to match the main text and because the dotless i does not work here.

Please make all indicator functions bold. The indicator function is a bold "1". If not possible, please use an alternative notation for the indicator function such as "Ind".

Please replace this N by the symbol for the "set of natural numbers" (The hollow N). I am not able to paste it here. If not possible, please replace this by: "... and each integer xi is the number of ..." (maintaining the "i" as a subscript of "x").

Please make these two fractions look equal, if possible. k/2 (together) is preferrable.

doi:10.1038/nmeth.3772

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 $G^s[i-(k/2)...i+(k/2)+1]$ was cleaved by the DNase I enzyme in comparison to its total occurrence in (1) DHSs (DHS sequence bias approach) and (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 models with $k < 6$, and the amount of information gained in models with $k > 6$ is 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 the raw DNase-seq signal ($\hat{x}_i^s$) and the bias score signal[15] ($\hat{b}_i^s$). 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 were 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) was obtained by applying

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

The corrected DNase-seq signal generated by equation (1) may include negative values. Because some posterior statistical analyses required a signal consisting only of positive values, we 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

Q4. If you copy and paste the link in the browser it works. For some reason beyond my understanding, the link does not work by clicking in the pdf file. I would like to keep the link as it is. However, if you really need a clickable working link you can separate the file name from the rest of the URL, which reads: "We obtained the footprint predictions (all.footprints.gz)<ref 4> for K562 cells (DU) from ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/supplementary/integration_data_jan2011/byDataType/footprints/jan2011/."

Q4

the parameters from the original publication: the flanking component length varied between 3 and 10 bp, and the central footprint–region length varied between 6 and 40 bp. Afterward the footprints were filtered by a false discovery rate (FDR) of 1%, which was estimated on the basis of the FS distribution in each cell type[4]. Finally, we considered only predictions that occurred within DNase-seq hotspots, evaluated using the method first described by Sabo *et al.*[37]. We obtained all hotspots generated by Stamatoyannopoulous's lab in ENCODE[1] for cell types GM12878 (wgEncodeEH000492; GSM736496 and GSM736620), H1-hESC (wgEncodeEH000496; GSM736582) and K562 (wgEncodeEH000484; GSM736629 and GSM736566). We refer to this framework as the Neph framework.

Boyle method. Boyle *et al.*[5] designed a segmentation approach 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 a 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 or increase in DHS signals around the peak-dip-peak region. Because no source code or software was provided, we used footprint predictions from Boyle *et al.*[5] available at http://fureylab.web.unc.edu/datasets/footprints/. We refer to this method as the Boyle method.

CENTIPEDE. CENTIPEDE is a site-centric approach that gathers experimental and genomic information around 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. We generated the experimental data input 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 the 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-scores.

CENTIPEDE software was obtained at http://centipede.uchicago.edu/ and executed to generate posterior probabilities of regions being bound by TFs. We had previously observed that CENTIPEDE is sensitive to certain parameters. Therefore, we defined CENTIPEDE parameterization with an extensive computational evaluation described by 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 TFBSs. It is based on a probabilistic classification approach to compute better log-posterior odds score than the ones observed with purely sequence-based approaches. We applied this method as described by Cuellar-Partida *et al.*[10]. We created a smoothed

Q5. The author moved the URL. Please update this URL to: http://tlbailey.bitbucket.org/supplementary_data/Cuellar2011/

Q5

the priors to the program FIMO[41] to obtain the predictions. We refer to this method as the Cuellar method.

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. On the basis of the observation that DNase-seq cuts from the double-hit protocol are strand-specific, Wellington considers only 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 obtained Wellington's source code from [illegible] default parameters were set to FDR = 0.01, −20, −30, footprint sizes varying between 6 and 40 bp with 1-bp steps, and a shoulder size (flanking regions) of 35 bp.

Maybe add "(PIQ)" here in the section header (before the dot).

Protein interaction quantification. PIQ is a site-centric method that uses a Gaussian process to model and smooth the footprint profiles around candidate MPBSs[11] (±100 bp). Active footprints are estimated with an expectation propagation algorithm. Finally, PIQ indicates the set of motifs for which footprint signals are distinguishable from noise to reduce the set of candidate TFs. We obtained PIQ implementation from http://piq.csail.mit.edu and e[illegible]h can be found in t[illegible]generated with the script pwmmatch.exact.r. The DNase-seq signal was created using the script bam2rdata.r. The footprints were detected with the script pertf.r.

Change to "Yardimci" to match main text.

Footprint mixture (FLR). Yard1mc1 *et al.*[12] proposed a site-centric method based on a mixture of multinomial models to detect active and 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 either initialized with 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 on the basis of the DNA sequences extracted within the same regions in which the cuts were retrieved. Method implementation was obtained from 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 a TFBS (PadLen) was set to the default value of 25 bp. Also, we used the expectation maximization to re-estimate background during training (argument Fixed set to FALSE). We refer to this method as the FLR method.

DNase2TF. DNase2TF is a segmentation approach based on a binomial *z*-score that evaluates the depletion of DNase-seq reads around the candidate footprints[7]. In its 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] proposed the segmentation method HINT as an extension of the 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 the HINT method starts with normalization of the DNase I cleavage signal using within– and between–data set normalizations. Then the slope of the normalized signals is evaluated to identify increases and decreases in the DNase-seq signal. Next an HMM is trained in a supervised manner (maximum likelihood) on the basis of a single manually annotated genomic region. To aid in 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 applied to the trained HMMs inside regions consisting of DHSs extended by 5,000 bp upstream and downstream. All parameters were set as described by Gusmao *et al.*[8].

We made two modifications to the method described by Gusmao *et al.*[8]. First, to perform a standardized comparison, we modified HINT to allow only DNase-seq data. The modified HMM model contained 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 before normalization steps. These modifications required retraining of the HMMs. For this, we used the same manual annotation described by Gusmao *et al.*[8]. The novel methods and trained models are available as a command-line tool at http://costalab.org/publications-2/hint-bc/.

BinDNase. BinDNase is a site-centric method based on logistic regression that is used to predict active and inactive MBBSs[13]. The algorithm starts with the base pair–resolution DNase-seq signal around MPBSs (±100 bp) 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 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 that requires TF ChIP-seq examples for training. We also point out that BinDNase did not execute successfully for 19 TFs in our evaluation data set (POU5F1, REST, RFX5, SP1, SP2, SRF, TCF12 and ZNF143 binding in H1-hESC cells; ARID3A, CTCF, IRF1, MEF2A, PU1, REST, RFX5, SP1, SP2, STAT2 and ZNF263 binding in K562 cells) given c[illegible]entation v[illegible]indnase/ a[illegible]

Please remove this comma. Also, the middle comma should separate "mi" and "ni" and not be a subscript of "mi".

[illegible] scheme termed the footprint score (FS), which is based on a scoring metric from the footprinting methodology proposed by Neph *et al.*[4]. The FS statistic is defined as

Remove this comma.

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

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

doi:10.1038/nmeth.3772

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

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

TC-Rank. The site-centric method referred to here as the tag count (TC) corresponds to the number of DNase I cleavage hits in a 200-bp window around predicted TFBSs as defined by He *et al.*[15]. This can be expressed 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, their respective methods (TC-Rank and FS-Rank) consist of attributing these quality scores to each MPBS and evaluating the performance at these ranked MPBSs. This observation also holds for the PWM-Rank method described below.

Evaluation. *MPBSs.* Method evaluation was performed with site-centric binding site statistics. For this we generated PWMs from PFMs by evaluating the information content of each position and performing background nucleotide frequency correction[42]. This was done 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 by Wilczynski *et al.*[45] with an 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 in a baseline method referred to here as PWM-Rank.

Method comparison. Methods were evaluated using a site-centric approach[10] that combined 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, and MPBSs without ChIP-seq evidence are considered false TFBSs. Every TF prediction that overlaps a true TFBS is considered a correct prediction (true positive), and every prediction that overlaps a false TFBS is considered an incorrect prediction (false positive). Therefore, true negatives and false negatives 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 the specificity of the method is decreased. The AUC metric was evaluated at 100%, 10% and 1% FPR. We also evaluated the AUPR. This metric is indicated for problems with imbalanced data sets (distinct numbers 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 and then ranked all nonpredicted sites by their tag count. To present a fair comparison, we also applied this approach 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 and 0.99 for the detection of footprints in all site-centric methods. We performed additional experiments to select the best threshold for each method (**Supplementary Fig. 8**).

Our TF ChIP-seq–based comparative experiments comprised the following three evaluation scenarios (He, benchmarking and comprehensive data sets). All evaluation statistics and details on method performance are available in **Supplementary Data Set 1**.

He data set. To replicate the analysis performed by He *et al.*[15], we analyzed DNase-seq data from cell types K562 (UW), LNCaP (DU) and m3134 (UW) for 36 TFs and evaluated the methods PWM, FS, TC, HINT, HINT-BC and HINT-BCN.

Benchmarking data set. For comparative analysis of several competing methods, we selected the two cell types with the highest numbers of TF ChIP-seq data sets evaluated in our study: K562 (DU), with 59 TFs, and H1hESC (DU), with 29 TFs. We were therefore able to make use of predictions provided by Gusmao *et al.*[8] and Boyle *et al.*[5], including evaluation of the PWM, Boyle, Cuellar, CENTIPEDE, HINT and Neph methods. For this data set, we estimated novel footprints for the FS, TC, HINT-BC, HINT-BNC, DNase2TF, PIQ, Wellington and FLR methods, which were not previously evaluated.

Comprehensive data set. Lastly, we compiled a comprehensive data set containing 233 [illegible] ing cellular backgroun[illegible] of 144 TF ChIP-seq and [illegible] 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.

Change to "Yardimci" to match the main text.

Expression-based evaluation (FLR-Exp). As shown by Yard1mc1 *et al.*[12], ChIP-seq evaluation of putative TFBSs may present biases regarding the fact that ChIP-seq data alone are not able to distinguish direct from indirect binding events. Consequently, we performed an evaluation procedure that combined MPBSs with differentially expressed genes from two cell types. The method evaluated the association of the quality of footprints overlapping particular motifs and the expression of the relevant 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 nonredundant PFMs from JASPAR in which the gene symbol was a perfect match with genes present in the array platform. This led us to 143 PFMs (**Supplementary Data Set 2b–d**). We applied genome-wide motif matching using these PFMs.

Next we evaluated the FLR[12] score, TC[15] and FS[15] for the footprints of each evaluated method that intersected with MPBSs of a particular motif. We considered only the footprints within DHSs that were common between the two cell types being evaluated, as described by Yard1mc1 *et al.*[12]. We expected that TFs expressed in cell type A would present higher values for 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 metric's 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 greater differences. As the KS score does not indicate the direction of the change in distribution, we obtained a signed version by multiplying the KS statistic by −1 in cases where the median of A was less than the median of B. We calculated the Spearman correlation between the signed KS test statistic and the expression fold change for each TF (**Supplementary Figs. 1** and **2**). Positive values indicated an association between expression of TFs and quality of footprint predictions. We call this correlation FLR-Exp. Results for FLR-Exp analysis are summarized in **Supplementary Data Set 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 or TF. As previously indicated by Sung *et al.*[7], TFs with shorter binding times characteristically have fewer DNase-seq cuts (protection) surrounding the binding site. More formally, the protection score for a set of MPBSs is defined as

$$\mathrm{PROT}_{\mathrm{MPBS}} = \sum_{i=1}^{N} \frac{\left(N_{\mathrm{R},i} - n_{\mathrm{C},i}\right) + \left(n_{\mathrm{L},i} - n_{\mathrm{C},i}\right)}{2N}$$

where MPBS = {MPBS_1, …, MPBS_N} is a set of binding sites for a given motif; $\mathrm{MPBS}_i = [m_i, n_i]$ is the genomic location of the ith binding site; and $n_{\mathrm{C,i}}$, $n_{\mathrm{L,i}}$ and $n_{\mathrm{R,i}}$ 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 between DNase-seq counts in the flanking region and DNase-seq counts within the MPBS. Positive values indicate protection in the flanking regions, whereas values close to zero and negative values indicate no protection. The protection score is similar to the FS[15]. The main difference is that the FS measures the ratio between reads in flanking and binding sites, whereas the protection score measures the difference. Finally, because we were interested in using the protection score as a measure of quality for a given TF and set of footprint predictions, we evaluated only MPBSs overlapping footprints for a given cell type. The DNase-seq count values had been previously corrected for DHS sequence bias and coverage differences. Results for protection scores are provided in **Supplementary Data Set 1**.

Statistical methods. We used the nonparametric Friedman-Nemenyi hypothesis test[48] to compare the AUC and AUPR of the methods regarding all data set combinations (TFs versus cell types). Such a test provides a rank of the methods as well as the statistical significance of the outperformance of a particular method. All correlations are based on Spearman values. All reported P values were corrected using the Benjamini-Hochberg method[49].

Code availability. Software, custom code, benchmarking data, DNase-seq sequence bias estimates and additional graphical results are available at http://costalab.org/publications-2/hint-bc/. HINT, HINT-BC and HINT-BCN software can be accessed directly through the Regulatory Genomics Toolbox website at http://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. & Furey, 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 Bioinformatics* **10**, 82 (2009).
46. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. 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. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

doi:10.1038/nmeth.3772

QUERY FORM

Nature Methods	
Manuscript ID	[Art. Id: 3772]
Author	
Editor	
Publisher	

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer queries by making the requisite corrections directly on the galley proof. It is also imperative that you include a typewritten list of all corrections and comments, as handwritten corrections sometimes cannot be read or are easily missed. Please verify receipt of proofs via e-mail

Query No.	*Nature of Query*
Q1	Please carefully check the spelling and numbering of all author names and affiliations.
Q2	Please check that all funders have been appropriately acknowledged and that all grant numbers are correct.
Q3	Please check that the Competing Financial Interests declaration is correct as stated. If you declare competing interests, please check the full text of the declaration (at the end of the main references section) for accuracy and completeness.
Q4	Please check. link not working
Q5	Url not working Please check.
Q6	Labels are overlapping. Please check.
	All queries were answered directly in the manuscript and are numbered accordingly. Furthermore, additional comments and corrections were made directly in the manuscript.