Dear Reviewers,

We would like to thank you for the thorough review of our paper and the constructive comments. We have included in the analysis requested transcription factors and DNase-seq experiments and expanded our evaluation with additional footprinting methods. Novel analysis further reinforces our initial findings that several state-of-the-art footprinting methods are not affected by cleavage bias. Moreover, our manuscript includes now an analysis of cleavage bias on TFs with short binding time. We observe that footprinting methods have poor predictive performance for these factors and that no clear footprint shape can be detected even after bias correction. We propose, therefore, a novel statistical measure (protection score) to computationally identify TFs with short binding time from footprint predictions and motif-predicted binding sites alone. The proposed score can successfully discriminate TFs with short and intermediate/long binding times and significantly correlates with the performance of footprint methods.

We extensively revised the main manuscript according to the Reviewers' comments. Changes in the supplementary material are marked in red.

Reviewer 1

"The analyses by Gusmao et al. are informative with the respect to comparison of computational footprinting methods, which was not systematically performed by He et al. This is an important clarification and is of general interest to the field."

We thank the Referee for his/her positive comments.

"However, I am not convinced that results from Gusmao et al. disagree with conclusions of He et al. Rather they expand some of the He et al. analyses by including additional methods."

"1. I disagree with the claim by Gusmao et al. that He et al. claim that "the simplest method for detection of active binding sites possible, outperforms computational footprinting". The authors only made observations for the methods that were evaluated."

We agree with the referee that our manuscript mainly expands results from He et al, 2014 by inclusions of new methods and data. Also, that some of the passages of our initial manuscript were too strong. We rephrased our manuscript to correct this.

However, we still think that our results disagree with one of the main conclusions from He et al. 2014. Their last paragraph of the discussion states: *"Our analysis of DNase-seq data and ChIP-seq data for 36 TFs showed that the efficiency of DNase I footprints in recovering TF binding sites was associated with the extent to which the observed cleavage pattern differs from the intrinsic cleavage bias."* Our cleavage bias analysis, which is based on 14 methods and 88 TFs, clearly indicate that this statement do not hold for several state-of-the-art footprinting methods (Fig. 1a).

"2. The authors demonstrate improved footprinting of six factors after DNase bias correction. However, no results are shown for AR and GR, which was given as a main example of uninformative footprints in He et al. Can the authors provide such analyses? Furthermore, it would be useful to compare corrected footprints for between He et al. and Gusmao et al. on the same set of 6+ factors (Fig. 5, SI)."

We now analyze additional DNase-seq experiments and the nuclear receptors AR, ER and GR (Supplementary Fig. 6 – page 20). We have also expanded our figures to include expected bias signals and deproteinized DNase-seq signals as in He et al. In all cases, deproteinized DNase-seq and expected bias signals indicate bias-associated signals around the binding sites. Corrected DNase-seq signals differ from the uncorrected signals, and we also observe a slight improvement of footprint shapes of ER. This is not the case for AR and GR, where the corrected DNase-seq signal does not show a clear footprint profile. Concerning our prediction analysis, the AUC for AR (R1881 treatment), ER (40 and 160 minutes after estradiol treatment) and GR (dexamethasone treatment) increases after bias correction (HINT-BC AUC - HINT AUC). These changes are in the top quartile of AUC improvement after bias correction.

A recent paper demonstrated that nuclear receptors have a short binding time and are likely to be artifacts of DNase-seq cleavage bias (Sung et al., 2014). Moreover, they show that footprints from transient TFs have a low DNase I protection surrounding the footprint. To further investigate this, we propose a statistics, which measures the protection surrounding footprints matching sequence-predicted binding sites (protection score; Supplementary methods). Indeed, the protection score on bias-corrected signals has negative values (no protection) around footprint predictions with binding sites of nuclear receptors AR, ER and GR (Fig. 2a, Supplementary Fig. 6 – page 20). Moreover, the protection score is able to separate TFs with short binding time (AR, ER and GR) from TFs with intermediate/long binding time (AP1/C-jun and CTCF) (Fig. 2a). We observe a significant Spearman correlation between the protection score and AUC values of all evaluated methods, i.e. R=0.19 (TC) and R=0.26 (HINT-BC).

Altogether, this indicates that computational methods for DNase-seq analysis have indeed poor performance on the TFs with short binding times. Moreover, our analysis shows that transitivity of TFs (as measured by the protection score) is a more relevant feature than DNase I cleavage bias for indicating the predictive performance of footprint methods. The protection score, which is estimated only on footprints and motif-predicted binding sites, can be used as a measure of quality of predicted footprints. These results are now discussed in the main manuscript.

"3. It is currently difficult to evaluate the results for a specific TF from the graph (Fig. 3, SI). Can the authors summarize the results of their AUC vs OBS analyses in a spreadsheet wtr to individual TFs and footprinting methods, so that the performance of methods can be evaluated for individual TFs?"

We have now included in the submission a table with the AUC values and all statistics used in our analysis (Supplementary Dataset 1).

“4. Can the authors comment on He et al.'s evaluation of 0500 and 0458 motifs? Do Gusmao et al. agree that these footprints are artifacts of DNaseI?”

Indeed these footprints are likely artifacts of DNase I cleavage bias. After correction of cleavage bias, the DNase-seq signal contains no clear footprint shape (Supplementary Fig. 7 – page 21). Next, we compared the overlap between footprints generated by HINT-BC and ‘Neph’ method (originally used to find these *de novo* motifs) and MPBSs for these *de novo* motifs. We observed that 24.99% (motif 0458) and 28.58% (motif 0500) of MPBSs are associated with a footprint predicted by ‘Neph’ method. In contrast, only 0.73% (motif 0458) and 1.71% (motif 0500) of MPBSs overlapped with HINT-BC footprints. This indicates that such motifs would not be detected on footprints predicted after bias correction.

“He et al. evaluated the effects of differential footprinting for predicting differences in TF binding between conditions and demonstrated high performance of this metric compared to footprinting score. It would be informative to see Gusmao and co-authors evaluate HINT and other methods for similar analyses.”

We agree that differential footprinting is a relevant problem. Note however that most data and methods evaluated here are not suitable for this problem. We think therefore this problem is out of scope for this manuscript.

Reviewer 2

"1) The authors are overstating the implications from He et al. The main point of that study was to raise concerns that simple methods that do not account for background, including some that had been used in high profile publications, could lead to false positive footprints (they do not suggest that no method would ever be able to identify footprints). In my opinion, this was shown unequivocally."

Indeed, some of the passages of our initial manuscript were too strong. We have improved our manuscript to clarify this issue. Note, however, that our results clearly disagree with particular text passages from He et al. 2014. Their last paragraph states: *"Our analysis of DNase-seq data and ChIP-seq data for 36 TFs showed that the efficiency of DNase I footprints in recovering TF binding sites was associated with the extent to which the observed cleavage pattern differs from the intrinsic cleavage bias."* Our analysis, which is based on 14 methods and 88 TFs, indicates that this statement do not hold for several state-of-the-art footprinting methods (Fig. 1a).

"More importantly, they authors have missed several studies that have already addressed this very question, published as independent papers rather than a short note."

We agree with the referee about the discussion of further studies and inclusion of competing methods on our analysis. Indeed, DNase-seq cleavage bias correction was already performed in Sung et. al, 2014 for 2- and 4-mers. Moreover, Yardımcı et al. used cleavage bias signals to build background models to detect inactive footprints. However, none of these methods were evaluated in regard to their sensitivity to cleavage bias or put into a large evaluation study as presented here. We have now extended our analysis to include these studies.

Our evaluation indicates that methods performing signal smoothing or that use 6-mer cleavage bias estimates are not affected by cleavage bias (Fig. 1a). From an overall performance aspect, we observe that all segmentation-based methods and PIQ outperform the TC approach (Supplementary Tables 2 and 3 – pages 23 and 24, respectively; Supplementary Fig. 5 – page 19).

"2) There is an issue with the estimation of the bias. Rather than explicitly using the bias parameters estimated on dechromatinized DNA by Crawford and Ohler for the Duke protocol, or earlier by Stamatoyannopoulos for the U Wash protocol, this is done using aligned reads inside DHSs. This is of course open chromatin, but not equivalent to naked DNA. This problem is reflected in supplementary figure 1, which shows the correlations of these 6-mer bias values in different datasets and protocols."

The DNase-seq signal from deproteinized/naked DNA is an alternative for estimating bias. Our initial strategy was based on the proposal from He et al., which estimated bias profiles inside DHSs. Theirs (and ours) analyses indicate similar DNase cleavage profiles using deproteinized DNA and bias estimates within DHSs (Supplementary Fig. 2, 6 and 7 – pages 16, 20 and 21, respectively). To further explore this topic, we evaluated the performance of

HINT with both bias correction strategies on experiments based on single- and double-hit protocols. We find that both strategies are better than no correction, but that the DHS-based correction (HINT-BC) is slightly superior to the deproteinized DNA approach (HINT-BCN) (Supplementary Fig. 5 – page 19 and Supplementary Table 3 – page 24). Moreover, a clustering analysis of the k-mer cleavage bias shows a few cases where there is low agreement between k-mers from deproteinized DNase-seq and regular DNase-seq. One possible reason is that small variations of the same protocol might introduce further cleavage bias. Moreover, we observe that DHS-estimated k-mers with high cleavage bias have a high CG content (Supplementary Fig. 4 – page 18). This was shown to be beneficial with regard to the recovery of the footprint profile on the CG-rich transcription factors (Fig. 1b and c and Supplementary Fig. 2 – page 16). Therefore, our analysis supports the use of reads inside cell-specific DHSs for bias correction.

"Crawford and Ohler showed that the bias in Wash vs Duke protocols is positively correlation (0.74), whereas here this appears to be much lower or not detectable."

The analysis from our original publication was based only on reads around DHSs, while the analysis from Crawford and Ohler (Yardımcı, et al, 2014) was based on deproteinized DNA experiments. Our current analysis includes now both strategies (see Supplementary Fig. 3 – page 17). We also observe a high Spearman correlation between cleavage bias from deproteinized DNA from K562 and MCF7 cells (R=0.99); and K562 and IMR90 cell (R=0.81), which is in agreement with Yardımcı, et al, 2014.

"-) The authors do not include the nuclear receptors in their studies (with a note that they are not represented in major PWM repositories). This is the "poster child" for factors that likely do not leave discernible footprints, and they need to include those."

Our analysis includes now AR, ER and GR factors as well as new DNase-seq experiments (Supplementary Fig. 6 – page 20). In all cases, deproteinized DNase-seq and expected bias signals indicate bias-associated signals around the binding sites. Corrected DNase-seq signals differ from the uncorrected signals, and we also observe a slight improvement of footprint shapes of ER. This is not the case for AR and GR, where the corrected DNase-seq signal does not show a clear footprint profile. Concerning our prediction analysis, the AUC for AR (R1881 treatment), ER (40 and 160 minutes after estradiol treatment) and GR (dexamethasone treatment) increases after bias correction (HINT-BC AUC - HINT AUC). These changes are in the top quartile of AUC improvement after bias correction.

A recent paper demonstrated that nuclear receptor factors have a short binding time and are likely to be artifacts of DNase-seq cleavage bias (Sung et al., 2014). Moreover, they show that footprints from transient TFs have a low DNase I protection surrounding the footprint. To further investigate this, we propose a statistics, which measures the protection surrounding footprints matching sequence-predicted binding sites (protection score; Supplementary Methods). Indeed, the protection score on bias-corrected signals has negative values (no protection) around footprint predictions with binding sites of nuclear receptors AR, ER and GR (Fig. 2a, Supplementary Fig. 6 – page 20). Moreover, the protection score is able to separate TFs with short binding time (AR, ER and GR) from intermediate/long binding time TFs (AP1/C-jun and CTCF) (Fig. 2a). We observe a significant Spearman correlation

between the protection score and AUC values of all evaluated methods, i.e. R=0.19 (TC) and R=0.26 (HINT-BC).

Altogether, this indicates that computational methods for DNase-seq analysis have indeed poor performance on the TFs with short binding times. Moreover, our analysis shows that transitivity of TFs (as measured by the protection score) is a more relevant feature than DNase I cleavage bias for indicating the predictive performance of footprint methods. The protection score, which is estimated only on footprints and motif-predicted binding sites, can be used as a measure of quality of predicted footprints. These results are now discussed in the main manuscript.

"-) In their analysis they find that some of the methods are not significantly influenced by bias … These broad differences between methods that use larger regions vs just sites is not clearly explained and phrased and should be revised."

We agree with the referee that these aspects are relevant indicators of how methods will cope with bias. Indeed, methods performing smoothing (or using large regions) as PIQ and Cuellar are not affected by cleavage bias. We have therefore included in the supplement a table describing the characteristics and performance of all evaluated competing methods (Supplementary Table 2 – page 23).

"Minor concerns: The labels in figure 1A appear to be wrong, they don't agree with the text. (Figure 1A is the same as supplementary figure 3, this seems a bit redundant)"

The labels from Fig. 1a have been corrected and duplicated figures in the supplement have been removed.

"The method, HINT, appears like a promising approach. With the size constraints imposed by the format, there is too little information how this is trained and whether and how it might differ to the original publication."

There are two main differences between HINT (described in Gusmao et al., 2014) and HINT-BC/HINT-BCN. First, we use bias-correction estimates previous to any pre-processing step. Two bias-correction strategies were performed: using bias estimates from reads inside DHSs regions (HINT-BC) and using bias estimated from deproteinized DNase-seq experiments (HINT-BCN). The second modification concerned adapting the original HINT model, which was originally based on using both ChIP-seq from histone and DNase-seq. This was achieved by removing histone-associated states from the hidden Markov model as well as histone signals from the emissions. These modifications also required retraining HINT, HINT-BC and HINT-BCN models using the same region/annotations reported in Gusmao, et al. 2014. These models are provided in our web resource and can be used to obtain footprint estimates from any DNase-seq experiments. We have expanded the description of HINT in the Supplementary methods with these points.

"Some results and comparisons would benefit from a more indepth discussion. For instance, the original Cuellar Partida is a prior derived from tag counts, and the comparably worse performance seems to result from a combination with FIMO (PWM scores). Similarly, "Neph" is compared against FS score (which is pretty much the same without optimized flank

scores) but performs much better. Insights into the flank optimization (which in the original Neph appears to be done manually/ad hoc) would certainly help others in the field."

Concerning Cuellar method, it is hard to evaluate the effect of the use of a distinct approach for PWM detection, as the method is intrinsically coupled to FIMO. We have already optimized FIMO parameters to make PWM predictions more comparable (Gusmao et al., 2014). The low performance of Cuellar method possibly stems from its simple modeling of DNase-seq signals. Note that in the current version we rank all footprints by TC reads, as this leads to prediction improvements of all site-centric methods. After this modification, Cuellar has a higher average AUC than TC.

Given its simplicity, the FS can be only seemed as a baseline method. Therefore, methods performing footprint flanking region optimization as Neph, Welligton and DNAse2TF clearly outperform FS. One method with particular good performance is DNAse2TF, which is only outperformed by HINT variants, did not require any special parameterization and was simple to execute. We have included these points in Supplementary Table 2 (page 23).

Reviewer 3

"In this correspondence the authors address the overall bias across all factors, but do not show how there method corrects for the bias for specific factors which have short binding times."

"For example, one of the key finding in He et al was that AR and p53 can not be accurately detected by DNase-seq footprinting. Does HINT-BS accurately detect AR or p53 footprints? With out this included it is hard to understand if correcting for bias will make a difference."

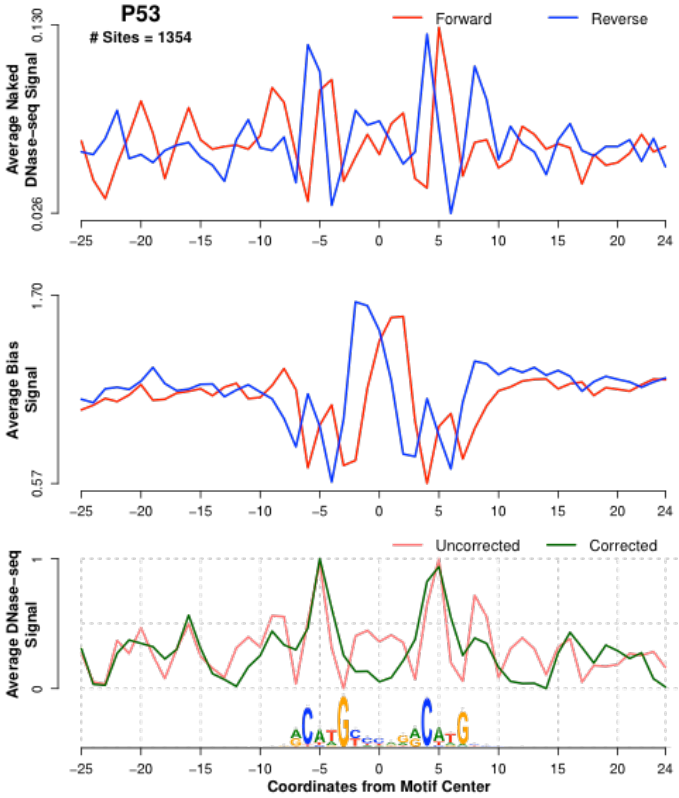
We include now in our analysis AR (R1881 treatment) as well as the nuclear factors ER (40 and 160 minutes after estradiol treatment) and GR (dexamethasone treatment) (Fig. 2, Supplementary Fig. 5 – page 19). Corrected DNase-seq signals differ from the uncorrected signals, and we also observe a slight improvement of footprint shapes of ER. This is not the case for AR and GR, where the corrected DNase-seq signal does not show a clear footprint profile. Concerning our prediction analysis, the AUC for AR, ER and GR increases after bias correction (HINT-BC AUC - HINT AUC). These changes are in the top quartile of AUC improvement after bias correction.

A recent paper demonstrated that nuclear receptor factors have a short binding time and are likely to be artifacts of DNase-seq cleavage bias (Sung et al., 2014). Moreover, they show that footprints from transient TFs have a low DNase I protection surrounding the footprint. To further investigate this, we propose a statistic, which measures the protection surrounding footprints matching sequence-predicted binding sites (protection score; Supplementary Methods). Indeed, the protection score on bias-corrected signals has negative values (no protection) around footprint predictions with binding sites of nuclear receptors AR, ER and GR (Fig. 2a, Supplementary Fig. 6 – page 20). Moreover, the protection score is able to separate TFs with short binding time (AR, ER and GR) from intermediate/long binding time TFs (AP1/C-jun and CTCF) (Fig. 2a). We observe a significant Spearman correlation between the protection score and AUC values of all evaluated methods, i.e. R=0.19 (TC) and R=0.26 (HINT-BC).

Regarding P53, we have noticed that there is no DNase-seq experiment on the same cell type used to perform P53 ChIP-seq (Saos-2). He et al. used DNase-seq from K562 cells for prediction of P53 binding sites. For that, only P53 binding sites that overlapped with K562 DHSs were analyzed. However, the use of distinct cellular background and the possible lack of P53 expression in K562 cells (doi:10.1038/sj.leu.2402647) put in question that particular analysis. Footprinting evaluation of K562 DNase-seq on Saos-2 P53 ChIP-seq confirms this, as both HINT-BC and TC have extremely low AUC values (0.21 and 0.12, respectively) and the P53 footprints have the 10^{th} lowest protection score (see Reply Letter Fig. 1). Clearly whenever no binding occurs, DNase-seq profiles around motif-predicted binding sites can only be artifacts of bias. Given that the cellular background of P53 ChIP-seq is the only one not matching any DNase-seq experiment, we have opted to not include P53 in our analysis.

Altogether, this indicates that computational methods for DNase-seq analysis have indeed poor performance on the TFs with short binding times. Moreover, our analysis shows that

transitivity of TFs measured by the protection score is a more relevant feature than DNase I cleavage bias for indicating the predictive performance of footprint methods. The protection score, which is estimated only on footprints and motif-predicted binding sites, can be used as a measure of quality of predicted footprints. These results are now discussed in the main manuscript.



Reply Letter Figure 1 - Average K562 DNase-seq signals around binding sites of P53 on cell type Saos-2 that overlapped with K562 hypersensitive sites. In the top panel, we show the strand-specific average DNase-seq signal on deproteinized DNA experiments (MCF-7 cell type); the middle panel shows the strand-specific estimated cleavage bias signal; and the bottom panels shows the (1) uncorrected – observed DNase-seq cleavage signal and (2) corrected – DNase-seq signal after the bias correction. Bottom panel signals were standardized to be in [0,1]. Below the graphs, we show the motif logo estimated on the DNA sequences of these regions.