Dear Reviewers,

We would like to thank you for the thorough reviews of our paper and the constructive comments. We have extensively expanded our analysis and manuscript to perform an extensive evaluation of DNase-seq computational footprint methods. This revision includes a methodology for the evaluation of footprints based solely on footprint quality scores and gene expression data. This methodology does not require ChIP-seq of transcription factors and represents an alternative to conventional evaluation methodologies. We have also included a new footprinting method in the analysis (BinDNase) and expanded our evaluation on DNase-seq experiments bias to address the Reviewers' concerns. Finally, we extensively rewrote the main manuscript to accommodate the new manuscript format. In particular, the manuscript includes a more detailed description of evaluated methods and their performance.

Reviewer 2

"*This paper by Gusmao et al is a short contribution that developed out of an earlier letter to the editor. First of all, it is a strong point that now 14 footprinting methods are compared, so this turned into an all-inclusive analysis. It is also positive that they have now extended their HINT method to include bias correction inferred from naked DNA too, not only DHSs.*

We would like to thank the Reviewer for the positive comments.

"According to equation three, the authors estimate DNase sequence bias for each 6-mer by calculating the ratio of observed DNase cleavages to k-mer abundance genome-wide (in DHSs or whole genome). This approach is similar to used in He et al. and Yardimci et al. and is an established way of measuring bias. Both regular DNase-seq experiments and naked DNase-seq experiments are used to estimate bias, and the results are similar in terms of AUC.

The approach from estimating bias from DHS appears problematic. It is simply not "DNase bias", as this is purely driven by the enzyme and estimated on naked DNA. In supplementary figure 3, they show a correlation matrix of the bias values inferred from DHSs vs naked DNA, both for single and double cut protocols. While the naked DNA datasets cluster together, the DHS-inferred bias values show less agreement within each other. There are two such datasets that don't agree with anything else, for instance (HepG2 and K562, both UW). Since the underlying enzyme is the same, the bias values inferred for different datasets should in principle have a certain degree of agreement. The absence of this leads one to question the chosen strategy."

We agree with the Reviewer that the wording used in our manuscript was imprecise. Indeed, the strategy based on deproteinized (naked DNA) experiments is to measure directly the "DNase-seq cleavage bias". The correction strategy used on DNase-seq reads inside DHSs of proteinized experiments will address not only DNase-seq cleavage bias, but also

fragmentation bias induced by the local chromatin conformation, CG content bias, among others. We call the latter now “DNase-seq experimental bias”. We have changed the text to make this point more clear.

To further investigate the issue with the “outlier” DNase-seq experiments in the clustering analysis, we have expanded our analysis to include all cells in ENCODE Tier 1 and Tier 2. We analyze the bias in 61 DNase-seq experiments in total. When doing this we have noticed that the aligned reads from the HepG2 (UW) and K562 (UW) DNase-seq provided by the ENCODE project had a one base pair offset, which caused their distinct cleavage bias estimates, thus forming the third cluster reported in the previous submission. After checking (and correcting) all DNase-seq experiments, our clustering analysis gave a much clearer picture (Fig. 2). There are two groups: one based on DNase-seq proteinized experiments from DU (single-hit) and another group with proteinized experiments from UW (double hit) and deproteinized experiments. Moreover, deproteinized experiments forms a small sub-cluster within the previous one. This analysis reinforces that distinct protocols have distinct “DNase-seq experimental bias”. It also shows that bias from the UW protocol is closer to the “DNase-seq cleavage bias”. Nevertheless, there are still moderate correlations between DNase-seq experiments generated using different protocols (average r = 0.39). All these points are included in the first paragraph of the section “Results – Impact of experimental bias” (main manuscript page 3).

“A final note on this issue is the argument that DHS-based bias estimation improves the footprint predictions for factors with GC-rich motifs. Since DHSs correspond to regulatory elements like promoters and enhancers, and since these elements are known to be GC-rich, it is of course expected that when the bias is estimated from DHSs, the k-mers that seem to be most associated with bias will be the GC-rich ones. This trend is clearly observed in supplementary figure 4. Note that here, the naked DNA datasets do not show this trend. So, the DHS-estimated bias correction is in fact to a certain extent a GC-bias correction. DNase cleavages in DHSs are driven by many different factors (nucleosome positioning, chromatin remodeling, GC content of promoter sites and TF binding events) and may correspond to a useful background model, but it is clearly *not* DNase bias. This should at the very least be clarified and contrasted to the approach taken by others.”

We agree that the CG bias, which is typically enriched in regulatory regions will also be captured by the “DNase-seq experimental bias” approach. For example, there is a clear association between CG content and predicted bias as indicated in Sup. Fig. 6 (formerly Sup. Fig. 4). To further investigate this, we evaluated the association between CG content of all motifs and the performance of HINT-BC, HINT-BCN and HINT for individual TFs (see Sup. Fig. 4b). While, motifs described in former Sup. Fig. 4 (currently Sup. Fig. 5) present high CG content, we observed no correlation between CG content of the motifs and the individual AUC of each method: HINT = -0.0144, HINT-BC = 0.0254, HINT-BCN = 0.0108 (p-value > 0.05). Furthermore, we observed no correlation between CG content of motifs and differences in AUC: HINT-BC - HINT-BCN = 0.0188, HINT-BC - HINT = 0.0724, HINT-BCN - HINT = 0.0644 (p-value > 0.05).

Concerning previous approaches, the “DNase-seq experimental bias” method is similar to the one used in He et. al, 2014 and DNase2TF. The only other method which measures DNase-seq cleavage bias is FLR (*Yardımcı* et al., 2014). FLR evaluates only regions with binding sites of a particular factor at a time. Therefore, it will access the CG bias on a factor-dependent manner. While both strategies are equally valid, the application of the TF specific FLR correction strategy to segmentation methods as DNase2TF or HINT is not trivial, as they work on a TF-independent manner.

In our view, a straightforward solution to evaluate the bias correction strategies is to contrast the performance of HINT with (and without) the two bias correction strategies: HINT-BC with “DNase-seq experimental bias” and HINT-BCN with “DNase-seq cleavage bias”. As show in Sup. Fig. 4a, HINT-BC has higher AUC values than HINT-BCN also for 226 out of 233 factors evaluated. These results indicates that performing bias correction is important and that “DNase-seq experimental bias” has an advantage over “DNase-seq cleavage bias” strategy. This advantage stems from the fact “DNase-seq experimental bias” captures a higher number of experimental artifacts than DNase-seq cleavage bias. We have expanded the results sections to include all of the previously discussed points (second paragraph of section “Results – Impact of experimental bias”; main manuscript pages 3 and 4).

“The title suggests that the study addresses cleavage bias; however, the focus lies on the discussion of predictive performance of various footprinting methods, and an elaboration of differences in different kinds of bias modeling or their impact is missing. It is better to replace the title with what the current version of the paper addresses.”

We have followed the Reviewer's and editor's advice to rewrite the manuscript and focused more on the evaluation methodologies. We now include a more detailed description of the methods (see Table 1; Sup. Table 1; and Extended Methods). We also expanded the repertoire of evaluated techniques by including a recently published footprinting method (BinDNase). Furthermore, we increased the number of evaluation statistics to accommodate more rigorous evaluation standards. Finally, we included a novel evaluation methodology which stems from the observations made in *Yardımcı* et al. (2014) about the inability of ChIP-seq to capture indirect binding events. These evaluation metrics are described below (this letter, page 4).

“The authors use AUC values at 10% FPR to compare methods. While this is a reasonable choice, related publications used full AUC values or those at 1% FPR to assess performance. Furthermore, CENTIPEDE, PIQ and FLR are turned into binary predictions (footprint vs. no footprint) before assessing performance. This is understandable as authors aim to compare both segmentation and site-centric methods; however using a singular AUC metric at a specific threshold may unfairly affect the performance of models that make non-binary predictions. Footprinting methods tend to result in high sensitivity at low FPR thresholds since the most obvious footprints tend to fall in high confidence ChIP-seq peaks; 10% FPR is too unrealistic and the authors should at least give the 1% values as well.”

We followed the request of the Reviewer and have extended our evaluation, which now includes the AUC at 1%, 10% and 100% FPR. We also included the area under the precision-recall curve (AUPR). This index is indicated for imbalanced data (very different numbers of positive and negative examples). We also evaluated the choice of the binarization strategy for site-centric methods (BinDNase, Centipede, Cuellar, FLR and PIQ), i.e. filtering out footprints with probabilities lower than (0.99, 0.95, 0.90, 0.85 and 0.80). In short, we observed that the probability threshold of 0.9 was best for all methods but BinDNase (best at 0.8; Sup. Fig. 7). We used the best cutoff threshold in the final evaluation analysis.

Concerning the use of AUC at 1% FPR and AUPR, we observed almost no change in the ranking of evaluated methods (Fig. 4). One clear finding is the fact that baseline methods – such as the tag count (TC) – have significantly lower mean AUC at 1% FPR and AUPR values than the state-of-the-art footprinting methods. This is an indication that TC has poor sensitivity. This also speaks against conclusions from He et. al. (2014), which were based on the AUC at 10% FPR. We included a complete description of these issues in the section “Results – Comparative analysis of footprinting methods” (main manuscript page 4).

“The most problematic point is that the authors use solely ChIP to evaluate footprints. Using ChIP-seq peaks as a sole gold standard is not adequate, since a subset of these miss footprints due to indirect binding (Neph et al, Yardimci et al. Sung et al.), are known to contain artifacts (Teytelman et al. , Park et al.), and do not have the resolution to discover footprints at nucleotide resolution. Therefore, successful bias modeling need not result in better ChIP prediction accuracy, and this type of benchmarking has caused a lot of confusion in the field. The authors need to find a different way to address this; they may for instance follow the suggestion of Yardimci et al to evaluate presence or absence of footprints in DHS, for cell lines that express resp. do not express the relevant transcription factor. (Additional analyses on "differential peaks" has also been requested by the other reviewer.)”

We understand the Reviewer's concern that ChIP-seq data have several limitations (indirect binding, low spatial accuracy, etc.), which might induce several false positives in the gold standard. We also need to realize that currently there is no current computational or experimental methodology to resolve these artifacts. Moreover, ChIP-seq based analysis are restricted to cells, where several ChIP-seq experiments from transcription factors are available. However, this is the gold standard used in all previous works proposing and evaluating computational footprinting methods. Furthermore, all methods are evaluated using the very same conditions over several ChIP-seq experiments. Finally, evaluations are based on relative performance, i.e. ranking of methods per TF, as seen in Gusmao et al. (2014) with the use of the Friedman-Nemenyi test. These strategies mitigate the impact of TF-specific artifacts.

Nevertheless, we do agree that “ChIP-seq free” evaluation methodologies are of great importance and would be a good complement to the conventional ChIP-seq based analysis. Therefore, we propose now a methodology for the evaluation of footprinting methods by expanding observations made by Yardımcı et al. (2014). In short, the authors compared the

footprint likelihood ratio (FLR) score around footprints in two cell types, where a particular factor was known to be expressed in only one of the cell types. They showed that the FLR score was significantly higher for the cell type with expression of the factor. The same observation did not hold when the tag count (TC) statistics of footprints were used instead of the FLR score.

We expanded the previous idea by evaluating if the difference of FLR score distribution for a number of TFs over a pair of cells types is proportional to the differences of the expression of these TFs in the same cells. In short, for a pair of cells we estimated the FLR score on footprints overlapping all motifs for a given motif database. For each TF, we estimated differences of FLR score distributions between two cells by a signed non-parametric Kolmogorov-Smirnoff (KS) statistic. Higher KS values indicate higher differences between the distributions. Finally, we measured the Spearman correlation between the KS statistic and fold change expression between a same cell pair over all evaluated factors. The Spearman correlation indicates how the expression changes *vs* FLR score agree. We observe very high correlation scores for most evaluated methods (median r = 0.79) and correlations greater than 0.9 for top performing methods (Fig. 1 and Sup. Fig. 1). We termed this correlation value as “FLR-Exp” and used it to rank the algorithms. Interestingly, final ranking differs in only a few positions with regard to the ranking indicated by the ChIP-seq based evaluation (r = 0.88; Fig. 4).

We also evaluated a similar procedure by using TC and FS scores in substitution to the FLR score. Our results with TC are in line with Yardımcı et al. (2014). in that TC scores have a poorer association with expression (median r = 0.35). Indeed, in many cases positive (negative) signed KS scores did not match positive (negative) expression fold change (see Sup. Fig. 2). The use of FS score resulted in slightly smaller median correlation values (r = 0.73) than the FLR scores. The final ranking of methods produced by the use of FS and FLR scores were also quite similar (r > 0.89) and included always the same top scoring methods as the ChIP-seq based analysis.

Given its best overall agreement with expression, we opted to use FLR score in our evaluation analysis. The new analysis was performed on all non-redundant motifs in the Jaspar database that matched with the gene symbol resulting from the expression analysis by the comparison of H1-hESC vs K562, H1-hESC *vs* GM12878 and GM12878 *vs* K562 cells. A clear advantage of this methodology, in comparison to ChIP-seq based analysis, is the fact that it requires only expression data and motifs. On the other hand, this metric does not allow the evaluation of the methods in a factor-specific manner, as required for analysis of cleavage bias or TF residence time. We added a full description of the new methodology to the main manuscript (Section “Results – Association of expression with footprint quality as evaluation measure”, page 3) and the Extended Methods (Section 1.4.3, page 13), which now includes both evaluation approaches: FLR-Exp and ChIP-seq based analyses.

“Some methods (Wellington, Cuellar-Partida prior) included in this paper are not affected by sequence bias in the way He et al. described (depletion of DNase-seq signal centered at TF binding site), or they combine many additional features in addition to the footprint (CENTIPEDE and the authors' own approach). Methods do

not just separate into site-centric and footprint-centric, but also into footprint-specific and multi-feature, and to discern the contribution of bias and footprints, a more careful evaluation is needed. At the current stage, the comparison without any elaboration of differences in the binding sites they discover, or which features are informative, fails to add novel insight into the issue other than showing the success of multi-feature methods to predict ChIP-seq signals."

Size restrictions of the previous manuscript formats did not allow us to include the discussion of such relevant aspects. We now include a more detailed discussion of method's characteristics and performance in the main manuscript (Section "Results – Computational genomic footprinting methods", page 2) and the Extended Methods (Section 1.3, pages 7-11). This is summarized in a characteristics table, which indicates the main feature of each method (Table 1).

"The authors discuss TF residence times as discussed by Sung et al. to explain lack of footprints. This is a reasonable hypothesis for explaining poor performance of footprinting for certain nuclear receptors but not yet an established fact at this point. ["Sung et al.7 showed that short-lived TFs display a lower DNase I cleavage protection pattern". Sung et al. proposed this hypothesis, but did not prove it or test it extensively.]

The observations in the paper are largely a confirmation of the model inspired by the Sung et al. results, and are in line with the observations in Yardimci et al that bias modelling still may not lead to successful predictions across all factors. It is valuable to the footprinting field that observations about nuclear receptors are consolidated in an effort to dispel the idea that footprinting methods can be used for all DNA binding proteins, but the authors should be clear about their specific contribution."

We have changed our statements in the text to meet the Reviewer's suggestions. We have expanded our discussion regarding the lack of accuracy of footprinting methods to predict footprints of particular TFs (main manuscript Section "Footprint predictions and transcription factor residence time", pages 4 and 5). Indeed, the protection score/binding time only s explains the low performance a subset of such TFs. Previously discussed problems of ChIP-seq TF evaluation, for example, could be the cause of such low predictive accuracy.

"Correction of Equation 3 in Supp. Material. (o * R) / (r * O)

Explanation of how OBS is calculated for each TF in Fig1A is missing."

We have corrected/improved these text passages.

"There are some changes in Figure 1 that should be explained to the reviewers. Even though the datasets used to generate the two versions are almost exactly the same (with the exception that a few TFs were added in the current version), Centipede did not have a significant negative correlation with bias previously and now it does. Also,

it seems to have much better accuracy now. The authors should explain this and make sure that they indeed get correct information from the compared predictors."

The differences are due to (i) the inclusion of novel TFs and (ii) that we used the tag count (TC) statistic for sorting footprints predicted by site-centric methods. The latter was performed because we noticed that the AUC values of all site-centric methods significantly improved in relation to the use of their respective footprint scores. Note that the TC score was already used by most segmentation-based methods from the initial submission version and we thought this would lead to a more fair comparison basis. These modifications were previously described in the supplement of the previous submission. To reinforce this point, we included an analysis in the main text and the Extended Methods, which evaluates the performance of strategies to rank footprints and refer to this in the main text (Sup. Fig. 7).

Reviewer 3

"The authors have addressed my concerns in the updated manuscript and have added protection score to address the concern on binding time."

We thank the Reviewer for all the insightful suggestions that helped improve the quality of the paper.

References

Gusmao, E.G. et al. Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. Bioinformatics 30(22), 3143-3151 (2014).

Yardımcı, G.G. et al. Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection. Nucleic Acids Research 42(19), 11865-11878 (2014).