Reply to referees: Detection of Active TFBSs with the Combination of DNase Hypersensitivity and Histone Modifications

Eduardo G. Gusmão, Christoph Dieterich, Martin Zenke and Ivan G. Costa

We would like to thank all three referees for their valuable considerations. We addressed all questions posed by the referees by improving the manuscript and performing additional analyses. You can find below detailed comments to all major requests. All minor corrections have been incorporated into the manuscript.

Reviewer 1

a. Major

1. *I would be much happier if all references to "ROC curve" were changed to "ROC-like curve" in the text and captions of the main and supplemental text. Even though you have now noted in one place that you don't use the standard definition, using the term "ROC" is misleading and confusing. Fortunately "AUC" works out mathematically to be the same for both ROC curves and your "ROC-like" curves, so that needn't be changed. Definitions are important.*

We changed all occurrences of "ROC curve" to "ROC-like curve" in the text and captions of the main and supplemental text. The term "AUC" remains unchanged.

2. *To talk about specificity or sensitivity of a method you need to fix one of them. Please clarify in Table 1 how you picked a single value of specificity (sensitivity) for each method when applying the Friedmann-Nemenyi test. Also, in the text, please explain what it means to say that "Cuellar presented higher sensitivity and lower specificity". I'm not sure that is even a meaningful statement. At all specificity levels in Fig 2, Cuellar seems to have lower sensitivity. So state in what cases (and in what sense) Cuellar gives higher sensitivity.*

All compared methods result in a number of footprints for a given parameterization. An exception is Centipede, which returns the posterior probability of a footprint to be active. In order to calculate its sensitivity and specificity we have considered its predictions with posterior probability ≥ 0.99 as suggested by the authors (Pique-Regi, R. *et al.*, 2011). The sensitivity and specificity were estimated from those predictions, which are marked with circles or squares in the ROC-like curves.

Given such definitions of sensitivity and specificity, the statement “Cuellar presented higher sensitivity and lower specificity” refers to the findings of the Friedman-Nemenyi test based on all 83 transcription factors, in which Cuellar method was ranked first according to its sensitivity but next to last according to its specificity (see Table 1 and Supplementary Table 1). In Fig. 2, Cuellar's sensitivity was higher than other methods' for C-jun, SIX5 and YY1 in cell type H1-hESC and only for YY1 in cell type K562 (green and yellow squares above all other squares/circles). This set of ROC-like curves is a small representative set. Please check the web supplement for all ROC-like curves.

We have improved the text in the manuscript to reinforce these points (Sections 3.3.2 and 4.3).

Reviewer 2

1. *The manuscript makes a fundamental assumption about active binding sites with regard to histone modification state around the binding sites. It is not clear what fraction of TFs follow this pattern. Are there specific families where this observation is not necessarily true?*

We have performed additional analyses to evaluate this. First, we evaluated the association between AUC values obtained by our method and the transcription factor classification recently proposed in Wingender, E. *et al.* (2013). As described in Supplementary Section 3.7, we could not find any statistically significant difference between the AUC of transcription factors grouped by classes. We also evaluated if the information content of the PWMs is associated to the AUC. Interestingly, the correlation between information content and the AUC is very close to zero, which indicates that PWM quality is not an important factor to our method. Therefore, we could not identify any clear factor-specific behavior.

Note however that the amount of ChIP-seq from transcription factors is still small for such analysis, i.e. only 3 TF classes had more than 12 factors. The accumulation of further TF, histone and DNase-seq data on new cellular contexts is still necessary for a comprehensive analysis of this relevant issue.

2. *On a related note, what is the false negative rate of the method? The training is based on manual annotation of a particular locus and therefore does not capture the binding sites that do not follow the histone modification pattern. One way to look at this is to analyze the DNase hotspots not annotated to contain footprints by the HMM method. Are there are specific motifs enriched in these peaks?*

In cell type K562, there are 1488 (1.33%) of DNase-seq hotspots not annotated to contain footprints by the DH-HMM (with three histones modifications). Segmentation-based methods Boyle and Neph missed respectively 78035 (69.66%) and 61365 (54.78%) of such hotspots (see Supplementary Table 13). We also observed that 97.65% and 90.52% hotspots which the DH-HMM failed to annotate, also did not contain footprints given Boyle and Neph methods, respectively. To get further insights about these regions, we plotted a histogram of the average DNase-seq read counts inside hotspots with and without footprints for either method (see Supplementary Fig. 11). This analysis shows that Boyle and Neph fail to annotate footprints on regions with lower DNase-seq read counts. Finally, we took the hotspots without DH-HMM predicted footprints and ran the MEME-ChIP tool in order to detect *de novo* motifs and TOMTOM to associate enriched motifs to all motifs in Jaspar (Machanick and Bailey, 2011). Only three enriched motifs were present in more than 10% of the hotspots (see Supplementary Table 14). The first motif was a G rich sequence associated to TFs as ZNF263, Pax4, KLF5, SP and EGR families. The second motif had a weak sequence and was unrelated to any previously known motif.

In short, DH-HMM has a higher coverage of hotspots than other segmentation-based methods and we could not detect any strong sequence signal on hotspots missed by DH-HMM. See Section 4.4 and Supplementary Section 3.8 for a detailed discussion.

3. *The results also show multiple footprints within a single peak – do each of these footprints correspond to the binding of a different factor or is this an artifact of the way the HMM is set up? This is important information for downstream analyses such as motif discovery.*

Our assumption and from related studies (Neph, S. *et al.*, 2012; Jankowski, A. *et al.*, 2013; Boyle, A. P. *et al.*, 2011) is that multiple footprints within a DNase-seq hotspot indicates combinatorial binding. We observe that 66.12% of the footprints obtained with DH-HMM (with three histone modifications) are supported by a TF ChIP-seq enriched region (peak) and that 45.12% of the ChIP-seq peaks are supported by a footprint (see Supplementary Table 11). Boyle (and Neph) footprints have a higher percentage of ChIP-seq coverage 78.93% (74.73%), but the footprints covered a lower number of ChIP-seq peaks 14.93% (28.39%). These statistics mainly reflect the higher specificity (and lower sensitivity) of both Boyle and Neph already captured in our main evaluation procedure (Section 4.3). We do not have a clear way to evaluate “unsupported” footprints, as they should arise from TFs not measured in the ChIP-seq experiments used in this study.

We performed some further analysis to have an overall picture of combinatorial footprint statistics. For these, we estimated the average number of footprints per hotspot for all segmentation-based methods. We divided hotspots according to their occurrence in proximal and distal regions. We also used results from a recent analysis performed on H1-hESC and K562, which identified regions with extremely high or low degrees of co-binding (HOT and LOT, respectively) (Yip, K. *et al.*, 2012). As shown in Supplementary Table 12, all methods tend to find several footprints per hotspot. There was a clear trend in finding more footprints in proximal vs. distal gene regions and HOT vs. LOT. DH-HMM tends to have a moderate number of predictions; with more footprints per hotspot than Boyle and fewer footprints than Neph. Moreover, both Boyle (0.63 footprints/hotspots) and Neph (0.11 footprints/hotspots) clearly underestimate the number of footprints in LOT regions, where at least one footprint should be available. See Supplementary Section 3.8 for a complete evaluation.

Reviewer 3

1. *While the footprints can be detected without knowledge of the TF, their interpretation does obviously require a known PWM. This is not clear in all cases.*

Indeed, we did not make clear statements regarding such issue in the paper. We added and emphasized sentences in Section 5 in order to make clear that the interpretation of footprints require proteins' binding sequence affinity information (i.e. PWM).

2. *Also, from a biological point of view, it is highly surprising that ES cells would not need separate parameterization (but maybe the relevant parameters are in fact rolled into the preprocessing of the data...). Its chromatin landscape is very different from differentiated cells (cf de Wit, Bouwman et al Nature 2013), and it might be worth to add a comment in this regard.*

In order to explore the different chromatin landscape between H1-hESC and K562, we split the analysis on the selection of histone modifications (see Section 4.1 and Supplementary Section 3.5) by cell type. In fact, we observe that the ranking of best histone combination model varies regarding different cell types. For instance, given the top seven methods (according to the Friedman ranking in Supplementary Table 8) for H1-hESC, five contained H2A.Z; while only three contained such histone variant for K562. Furthermore, the model consisting of the single histone variant H2A.Z was at the bottom of the ranking in K562, while the individual model with H3K27ac had the worst AUC in H1-hESC. These observations fit with cellular context specific functions of such histone marks. For example, H2A.Z plays an important role in differentiation in Embryonic Stem Cells (ESCs) (Subramanian, V. *et al.*, 2013; Hu, G. *et al.*, 2012). On the other hand, H3K27ac (together with H3K4me1) is known to be associated to poised (and inactive) enhancers in stem cells (Rada-Iglesias, A. *et al.*, 2010).

Note however that most of the above mentioned differences are not statistically significant and that most three-histone models are equivalent in all scenarios (see Supplementary Tables 9 and 10). Possibly, the combination of several histones alleviates any potential "cell-specific" effect in the predictions. Indeed, our previous analysis on the impact on performance when using different cell types for training/evaluating (see Section 4.2 and Supplementary Section 3.6) shows that our model is robust enough to account for such variability and can be seen as an advantage of our method. We have included new results and expanded the discussion in Section 4.1 and Supplementary Section 3.5.

In summary, we successfully addressed all points made by the reviewers and very much hope that the paper can now be accepted in *Bioinformatics.* If you have any further question, please let us know.

References

Boyle, A. P. *et al.* (2011). High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome Research*, **21**(3), 456–464.

Hu, G. *et al.* (2012). H2A.z facilitates access of active and repressive complexes to chromatin in embryonic stem cell Self-Renewal and differentiation. *Cell Stem Cell*.

Jankowski, A. *et al.* (2013). Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers. *Genome Research*.

Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**(12), 1696–1697.

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

Pique-Regi, R. *et al.* (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. *Genome Research*, **21**(3), 447–455.

Rada-Iglesias, A. *et al.* (2010). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, **470**, 279–283+.

Subramanian, V. *et al.* (2013). H2a.z acidic patch couples chromatin dynamics to regulation of gene expression programs during esc differentiation. *PLoS Genet*, **9**(8), e1003725.

Wingender, E. *et al.* (2013). TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Research*, **41**(D1), D165–D170.

Yip, K. *et al.* (2012). Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biology*, **13**(9), R48+.