

Supplement to “Detection of Active Transcription Factor Binding Sites with the Combination of DNase Hypersensitivity and Histone Modifications”

Eduardo G. Gusmão¹, Christoph Dieterich², Martin Zenke^{3,4} and Ivan G. Costa^{1,5,6,*}

¹IZKF Computational Biology Research Group, Institute for Biomedical Engineering, RWTH Aachen University Medical School, Germany.

²Computational RNA Biology and Ageing, Max Planck Institute for Biology of Ageing, Germany.

³Department of Cell Biology, Institute for Biomedical Engineering, RWTH Aachen University Medical School, Germany.

⁴Helmholtz Institute for Biomedical Engineering, RWTH Aachen University, Germany.

⁵Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Germany.

⁶Center of Informatics, Federal University of Pernambuco, Brazil.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXXX

1 RELATED WORK

Recent studies have attempted to improve the active transcription factor binding site (TFBS) detection using data that reflect chromatin state. We have categorized them in two classes: segmentation-based methods and site-centric methods. The former segments the genome in different regions, including likely TFBSs, based on cell-specific experimental data. The second uses sequence information to identify TFBSs and then filters these locations for active sites based on experimental data. In a higher level, each method has its peculiarities and may address a particular problem which stems from a specific methodological framework. In addition, each method may have been evaluated with a different gold standard. Nevertheless, in this study we have performed a comprehensive comparison including both segmentation-based and site-centric methods. Following, we introduce the main competing methods.

In Won, K. J. *et al.* (2010), authors used data from the histone modifications H3K4me1 and H3K4me3 to detect, respectively, active promoter and enhancer regions. They propose the use of three state hidden Markov models (HMMs) to capture dip regions of the characteristic peak-dip-peak shapes of histone modifications indicative of active **promoters**/enhancers. An HMM was trained for each histone modification on regions with high number of reads obtained with chromatin immunoprecipitation followed by sequencing (ChIP-seq) and high scoring TFBSs and combined in a mixture model. They could demonstrate on data from embryonic stem cells that their predictions were superior to any of the sequence based TFBS methods.

The advent of DNase-seq technique, which allows genome-wide detection of open chromatin regions with high resolution,

allowed further improvements (Boyle, A. P. *et al.*, 2011; Pique-Regi, R. *et al.*, 2011). In Boyle, A. P. *et al.* (2011), an HMM was developed to detect footprints in specific DNase I digestion patterns derived from DNase hypersensitivity (DHS) data. Note that the footprint definition is analogous to finding the dip in a peak-dip-peak previously explored by Won, K. J. *et al.* (2010), only the sizes of the peaks and dips are quite distinct between DHS and histone modification profiles. To cope with the higher resolution of the DHS signals, authors used a Savitzky-Golay filter to reduce noise and to estimate the slope of the DHS signal. The prediction task required a five state HMM, with specific states to identify the decrease/increase of DHS signals around the peak-dip-peak region. The HMM was trained on a supervised approach from a single annotated region.

Furthermore, Neph, S. *et al.* (2012) used an improved and conceptually simplified segmentation-based method originally proposed in Hesselberth, J. R. *et al.* (2009). Briefly, they use a sliding window approach in order to find regions (6–40 bp) with low DHS counts between regions (3–10 bp) with high DHS counts. A footprint occupancy score (FOS) is evaluated and used to determine the most significant putative TFBSs. As in Boyle, A. P. *et al.* (2011), the authors use only DHS data.

A quite distinct methodological approach, termed Centipede, was proposed in Pique-Regi, R. *et al.* (2011). Briefly, Centipede consists of scanning the genome for putative TFBSs, gathering experimental and genomic information around those regions and using an unsupervised Bayesian approach to label each retrieved site as ‘bound’ or ‘unbound’. The experimental and genomic data used included DHS and histone modifications ChIP-seq, motif matching bit-score, sequence conservation and distance to the nearest transcription start site (TSS). They used DHS as the main source of information for their predictions and found that histone modifications did not have a significant improvement in TFBS

*to whom correspondence should be addressed

detection for most transcription factors (TFs). Note however, that they use a statistic that summarized the histone occupancy around the site of interest, therefore not being able to detect the peak-dip-peak patterns characteristics of this data (Won, K. J. *et al.*, 2010).

Lastly, Cuellar-Partida, G. *et al.* (2012) proposed a technique to include DHS and histone modification data as prior probability in the detection of active TFBSs on a probabilistic classification approach. This study was also the first comparing approaches, where they showed that the accuracy of Centipede was mostly superior to their approach.

Moreover, the search of *de novo* motifs can be greatly enhanced by looking at more specific TFBSs, i.e. footprints closer and not much broader than the actual TFBSs. To support this claim we mention two recent studies. In Kulakovskiy, I. V. *et al.* (2009), the authors used footprints derived from a curation of 201 non-redundant DNase I footprinting studies in *Drosophila* Bergman, C. M. *et al.* (2005) to perform motif discovery. Such analysis allowed the automatic creation of 41 motifs for *Drosophila*, which were shown to have greater sensitivity than previous models. In Neph, S. *et al.* (2012) footprints derived from DHS were used for *de novo* motif finding. According to the authors, despite the efforts devoted to identify cognate recognition sequences of DNA-binding proteins, high quality motifs are only available for a small fraction of the TFs with predicted sequence-specific DNA binding domain. Their analysis on DHS regions found striking 289 footprint-derived motifs that were absent from major databases.

2 DATASETS

2.1 ChIP-seq, DNase-seq and PWMs

We present in Tables 22– 27 a summary of all the data used in this study, including the cell types HeLa-S3 and HepG2, which were used only in the empirical analysis described in Section 3.6. In Table 22 we present all DNase-seq and histone modification ChIP-seq experimental data used as input for our (and competing) methods. In this table we report the source lab, UCSC accession number, GEO accession number (if available) and total number of mapped reads used to create the genomic signal for each data track and cell type. All data presented in this table was obtained in the ENCODE repository (ENCODE Project Consortium, 2012).

The Tables 23– 27 exhibit all data used to create de validation datasets. Each table contains information on TFs, ChIP-seq and position weight matrices (PWMs) for one of the cell types. Each row contains the source lab and UCSC accession number for the TF ChIP-seq, as well as the repository and ID number of the PWM used to predict true and false motif-predicted binding sites (MPBSs) for that particular ChIP-seq data. All ChIP-seq data presented in these tables were obtained in the ENCODE repository (ENCODE Project Consortium, 2012) and all PWMs were obtained in the repositories Jaspar (Mathelier, A. *et al.*, 2014), Transfac (Matys, V. *et al.*, 2006) and Uniprobe (Robasky and Bulyk, 2011).

2.2 Gold Standard

Our gold standard was created based on checking overlapping status of MPBSs (sequence-based predictions obtained with motif matching) with cell-specific experimental evidence of binding

(ChIP-seq peaks, i.e. enriched regions). We have observed that previous work used distinct criteria for definition of thresholds for motif matching and detection of ChIP-seq peaks (Pique-Regi, R. *et al.*, 2011; Boyle, A. P. *et al.*, 2011; Cuellar-Partida, G. *et al.*, 2012; Whittington, T. *et al.*, 2009). Next, we briefly describe the main points concerning the choice of the parameters necessary in order to create a gold standard to evaluate the methods.

We obtained all TF ChIP-seq data from ENCODE (ENCODE Project Consortium, 2012). We downloaded the TF enriched regions (peaks) uniformly processed by ENCODE's Analysis Working Group (AWG).

Concerning motif matching, we evaluated the use of a fixed bit-score threshold of $\log_2(10^4) \approx 13.2877$ as proposed in Pique-Regi, R. *et al.* (2011) and a False Positive Rate (FPR)-based criteria (10^{-4}) (Wilczynski, B. *et al.*, 2009). The latter uses an approach based on dynamic programming to detect significant TFBSs. In this scenario, a different threshold is calculated for each motif by defining the bit-score that corresponds to a specific motif matching FPR-based threshold in the distribution of scores of that TF's PWM.

The Tables 17– 21 show statistics on the number of MPBSs, ChIP-seq peaks and combinations of both, for all cell types used in this study. These tables present data regarding the *p*-value selection criterion for peak detection and the two motif matching cutoffs mentioned. The bit-score of 13.2877 (always in the top line for each TF) correspond to Pique-Regi, R. *et al.* (2011) motif matching cutoff criterion; while the varying bit-score at each TF's bottom line correspond to the FPR-based motif matching cutoff. In all cases, the FPR-based selection criterion results in greater proportion of ChIP-seq peaks with an underlying MPBS than the fixed bit-score approach. For instance, only 17.54% of the GABPA peaks in cell type K562 contain overlapping MPBSs selected using bit-score approach, while 37.61% of these peaks contain overlapping MPBSs selected using the FPR-based selection. Therefore, we preferred to use the FPR-based approach, as the fixed bit-score approach represented overly conservative predictions (on average, only 27.71% of the peaks had their corresponding motifs).

3 HMM EXPERIMENTAL DESIGN

All parameter selection experiments described in this section were based on data from cell types H1-hESC and K562; and histone modifications H3K4me1 and H4K4me3 unless otherwise stated. All experiments were performed only on chromosome 1, which was removed from all further analyses.

3.1 HMM Inference

All HMM models were trained in 10,000 bp randomly selected regions. The models based on histone modifications H3K4me3, H3K9ac, H3K27ac and H2A.Z were trained in the region spanning from 211,428,000 bp to 211,438,000 bp in chromosome 1. This region encloses the promoter of the gene RCOR3, which is expressed in all cell types analyzed. The H3K4me1-based models were trained in the region spanning from 26,942,000 bp to 26,952,000 bp, also in chromosome 1. An inspection on ENCODE tracks on the genome browser indicates that this region is distal to any known gene and did not display any expression levels in the cell types analyzed.

Here we show an example of a complete set of HMM parameters, regarding the model trained with DHS + H3K4me3 using data from the H1-hESC cell type. The Table 2 represents the transition matrix. Each number represents the probability of performing a transition from the HMM state in the first column of the entry's row to the HMM state in the first row of the entry's column. The Table 3 exhibits the emission distribution mean values. It contains the mean in which each signal type (represented in the columns) assumes at each state (represented in the rows). Finally, the Table 4 shows all covariance matrices from the emission distributions. The full covariance matrix is depicted for each state, in which rows and columns are sorted by the input signals: DNase normalized, DNase slope, H3K4me3 normalized and H3K4me3 slope.

3.2 Choices Regarding Scaling Methodology

Although the first normalization step (within-dataset) is always used in a local fashion (Boyle, A. P. *et al.*, 2011), the second normalization step, i.e. the scaling procedure, can be performed using either local or global values of the standard deviation and percentile. The local method is depicted in Eq. 2 (main document), using the same windowing as the normalization procedure. The global method consists on using a single estimate of standard deviation and percentile for each cell type.

We have performed an empirical analysis to drive the choice of the scaling methodology and to test different percentile cutoffs. We created receiver operating characteristic (ROC-like) curves (with true negative rate (specificity) on x-axis, instead of the traditional false positive rate) and calculated the area under the ROC-like curves (AUCs) using our model with signals generated with either global or local methodologies combined with either the 96th, 98th or 99th percentiles.

We have created boxplots with the distribution of the differences between the local and global scaling approaches (Fig. 1A) and between the different percentiles tested (Fig. 1B). The distributions show a slight advantage for the local approach and for higher percentile values. Concerning the choice of percentile, both 98th and 99th percentiles are superior to the 96th percentile. We chose to use 98th percentile as it presents a more lenient criteria.

Loosely speaking, using the 98th percentile means that only the top 2% values will have a value greater than 0.5, since the scaling follows a logistic function. This is supported by estimates of the average coverage of DHS sites or regions enriched with histone marks over multiple cell types in human genome (ENCODE Project Consortium, 2012) (see Table 5). See Fig. 2 for example of normalized and slope signals using the selected scaling parameters.

3.3 Alternative HMM Topologies

The individual pattern of DHS around active TFBSs generally follows the peak-dip-peak trend when we consider aligned reads from both strands. This follows directly from the DHS protocol, wherein the DNase I enzyme nicks the DNA in loci that can be reached (Crawford, G. E. *et al.*, 2006; Song and Crawford, 2010; Boyle, A. P. *et al.*, 2011). However, although the average histone modifications trend generally presents a very clear (close to symmetric) peak-dip-peak pattern, there is inherent heterogeneity within individual *loci* regarding signal magnitude, asymmetry and implicit strand orientation (Kundaje, A. *et al.*, 2012; ENCODE Project Consortium, 2012).

We have therefore evaluated two additional HMM topologies depicted in Fig. 3. The model M2 is an extension of the original model (here denoted as M1) to account for the histone modification signal asymmetry, i.e. that some DHS sites have very small signals of active histone modifications on its downstream or upstream regions. For such, two additional transitions were added (shown in red) in order to allow the DNase level states to be visited when there are no histone modification peaks before or after DHS peaks. The new transition probabilities were estimated taking into account the proportions of asymmetrical peaks reported in Kundaje, A. *et al.* (2012). The model M3 is a simplification of M1, which performs the predictions of footprints without the slope signal. In M3, the UP, TOP and DOWN states from M1 are compressed into one state – HIGH – which recognizes high levels of DHS (DNase level state) or high levels of histone modifications (histone level state). Consequently, the HMM needs only the normalized signal and becomes bivariate (DHS and histone modifications normalized signals).

We tested these three models and obtained ROC-like curves for all combinations of cell types, histones and TFs tested. The Fig. 4 shows the distribution of the AUC differences between these models. We observe a slight advantage for M1 when compared to M2 (paired Wilcoxon-Mann-Whitney p -value = 2.07×10^{-20}) and a clear advantage of M1 and M2 over M3 (paired Wilcoxon-Mann-Whitney p -value = 4.78×10^{-58} and 6.35×10^{-37} , respectively).

The possible reason for the good results of M1 in relation to M2 is the fact that the normalization methodology emphasizes even little increases in histone levels leveraging the asymmetry issue (see Fig. 5). Furthermore, the poor performance of M3 in comparison to M1 and M2 indicates that even with more complex models (4 vs. 2 variables and 8 vs. 4 states, respectively) the slope signal and the additional states are crucial in the accurate delineation of the footprints.

3.4 Analysis of HMM decoding algorithm

In addition to the Viterbi algorithm implementation, we have tested the usage of posterior decoding (Rabiner, 1989) in order to generate our footprint predictions. In this case, we assume to be footprints contiguous genomic coordinates j with

$$P(q_j = FP | \mathbf{X}) > P(q_j = u | \mathbf{X}) \quad \forall \quad u \neq FP, \quad (1)$$

where \mathbf{X} is the matrix containing the observations (genomic signal) and FP represents the FOOTPRINT state.

We show in Fig. 6A the distribution of the pairwise AUC difference between predictions generated using the Viterbi method and posterior decoding. We are able to observe that there is a slight advantage in using the Viterbi method. Despite the lower AUCs for the posterior decoding, one advantage of such technique would be to use the HMM posterior probabilities as priors during operations such as motif matching (similarly as in Cuellar-Partida, G. *et al.* (2012)). However, as can be seen in Fig. 6B, the posterior probability of the HMM being in the FOOTPRINT state changes drastically from 0 to 1. Such “spiky” distribution would lead to results equivalent to the region filtering approach.

3.5 Analysis of alternative histone modifications

We have performed an empirical test on the predictive power of different histone modifications. All HMM models receive as input a

DHS signal and one histone signal, which in our test varied among the modifications H3K4me1, H3K4me3, H3K9ac, H3K27ac and the variant H2A.Z. All these histones signals are associated to active regulatory regions and are frequently measured in ChIP-seq studies. We also evaluated the combination of all pairs and triples of histone signals by simply merging all predicted sites, i.e. performing a union step and merging all predictions that overlapped. This pairwise combination generates 20 additional prediction sets (10 pairs and 10 triples). Note that extending to further combinations would deviate from one of the main goals, which is to create a consistent regulatory map with few genome-wide assays. We evaluated all the 25 combinations to all cell types and TFs tested. **In this analysis, we combined data from cell types H1-hESC and K562.**

The Fig. 7 presents the distribution of AUCs for all histone modification models tested. We can observe that most methods present the region between the first and third quartiles approximately between 80% and 95%. In order to test the statistical relevance of these differences, we performed a Friedman-Nemenyi test. The Table 6 shows the histone model ranking in decreasing order and their respective Friedman ranking. The Table 7 exhibits the full hypothesis test results, providing information on which models significantly outperformed others.

Overall, histone triples and pairs significantly outperform single histone models. Although a few histone triples significantly outperformed some histone pairs, this result is not as clear as the comparison of these models with individual histone modifications. We chose to report, in the main text, the results for the models consisting of the best histone triple (H3K4me1+H3K4me3+H3K9ac – which we denoted as DH-HMM(3)) and the best histone pair (H3K4me1+H3K4me3 – which we denoted as DH-HMM(2)). Note however that several other combinations would perform similarly well.

In addition, in order to explore the fact that H1-hESC and K562 cell types have different chromatin landscapes (de Wit, E. *et al.*, 2013), we have split the analysis presented in this section by cell type. The Friedman-Nemenyi results can be seen in Tables 8–10.

Interestingly, there are differences on the ranking between H1-hESC and K562. For instance, given the top seven methods (according to the Friedman ranking) for H1-hESC, five contained H2A.Z; while 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 AUCs in H1-hESC. One explanation for this is the fact that some histone modifications have distinct roles depending of the cellular context. 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 enhancers but not active sites in stem cells Rada-Iglesias, A. *et al.* (2010).

Note, however, that the difference in the individual Friedman ranking between these two cell types does not result in significant performance variability. The Friedman-Nemenyi results show that most three-histone models do not significantly outperform other three-histone models. The same behavior is true when we consider models with histone pairs or singles. This is in accordance with the Friedman-Nemenyi results for both cell types combined.

3.6 Application of the DH-HMM models on Different Cell Types

The annotation of a certain region with the HMM states is laborious. Consequently, it would be interesting to observe the performance of models trained with data from distinct cell types. For such, we have expanded our set of experiments two new cell types (HeLa-S3 and HepG2) each containing 20 and 21 TFs.

In order to test the previous claim, we have performed a paired Wilcoxon-Mann-Whitney test to compare the distribution of the AUCs for a given cell type with all four models. In particular, we are interested to observe if there is a significant decrease in AUC when the model was not trained with the data at hand. The Fig. 8 shows the results for all models applied to all cell types tested. Each set of four boxplots represent one of the four models, which was applied to the signal generated from the cell type labeled on the bottom of the set. Statistical significance on the pairwise difference between these distributions is represented by the three-star system.

We can observe in Fig. 8 that only in one out of twelve cases the AUC levels are significantly different (p -value ≤ 0.05). This corresponds to the HepG2 model, when used to generate footprints in the same cell type and in K562 cell type. These results suggest that our signal processing and models are able to robustly perform predictions over different cell types. Consequently, a simple application of the models already stored in our software tool is sufficient to generate accurate predictions.

3.7 TF-oriented Analysis of AUC Results

Although the epigenetic grammar around TFBSs is well-characterized for a number of factors ENCODE Project Consortium (2012); Neph, S. *et al.* (2012), it is not known to which extent such grammar applies to the great variety of human TF binding properties.

We have performed two analyses regarding this issue. The first analysis searches for correlations between our method's performance and features that describe TF binding affinity. In order to test for associations between these paired samples we used Pearson's product moment correlation coefficient. The Fig 9 shows a scatterplot between the PWM's information content (PWM's IC) and our method's AUC (when using three histone modifications). We have observed no correlation between these two features (correlation coefficient ≈ 0.0204).

In the second analysis we tested our method's AUC between different TF classes. For that, we obtained the TF classification (TFClass) from Wingender, E. *et al.* (2013). TFClass' scheme categorizes TFs based on the characteristics of their DNA-binding domains. It comprises six levels (in decreasing order of magnitude: superclasses, classes, families, subfamilies, genera and factor species). We have observed that the 'class' level would represent a good balance between the number of TFs we used and information about the TFs' binding characteristics. All TFs used in this study were categorized with exception of one (NRF1), which fell into the category of 'Yet Undefined DNA-binding Domains'. We present in Fig. 10 the distribution of our method's AUC given the TF classes. We performed a pairwise two-sided t -test between the classes with a considerable number of TFs (C2H2 zinc finger factors, basic leucine zipper factors, basic helixloophelix factors (bHLH) and tryptophan cluster factors). This test showed no p -value ≤ 0.1 .

The analyses we performed showed no correlation between our method's performance, i.e. the ability to recognize the epigenetic grammar for TF binding and many TF features. In this test comprising 83 TFs, our method seems not to be affected by characteristics of the TF's DNA binding. Note however that most classes had very few TFs associated (3 classes with more than 12 TFs). A larger number of TF ChIP-seq is required for a more thoroughly investigation.

3.8 Statistics on Footprints and DHS regions

We performed further analyses to investigate the coverage of footprints in DHS regions. Note that this study can be performed only on segmentation-based methods, since site-centric methods do not provide genome-wide footprint predictions (i.e. their predictions are already based on known protein-DNA binding affinity information).

It is common to encounter multiple footprints at a single DHS region (hotspot). Our main assumption (and from recent studies (Neph, S. *et al.*, 2012; Jankowski, A. *et al.*, 2013; Boyle, A. P. *et al.*, 2011)) is that this indicates combinatorial binding. First, we calculated the overlap between the footprint predictions from the three main segmentation-based methods and the combination of all ChIP-seq peaks used in this study for cell type K562 (see Table 11). We observed that 66.12% of the footprints obtained with DH-HMM (with three histones) are supported by a TF ChIP-seq enriched region (peak) and that 45.12% of the ChIP-seq peaks are supported by a footprint. It is not possible to evaluate the validity of "unsupported" footprints (i.e. footprint predictions without overlap with our set of ChIP-seq peaks), as they can arise from TFs not measured in the ChIP-seq experiments used. However, we point to the fact that our method covered more ChIP-seq regions than Boyle and Neph method (respectively, 14.93% and 28.39%), which reflect the higher sensitivity of our method.

Moreover, we can perform some further analysis to give an overall picture of combinatorial binding statistics. For these, we estimated the average number of footprints per DHS region for all segmentation-based methods in cell type K562. We divided DHS regions according to their occurrence in proximal and distal regions. In this analysis, promoter regions were considered as [TSS-1500,TSS+100] where the first point in the interval is always upstream. We also used results from Yip, K. *et al.* (2012), which identified regions with extremely high or low degrees of co-binding (HOT and LOT, respectively) based on the same TF ChIP-seq data used in this study. As shown in Table 12, all methods tend to find several footprints per DHS regions. There was a clear trend in finding more footprints in gene proximal vs. distal regions and HOT vs. LOT. This fits the definition of HOT and LOT regions and the expectation that more factors bind on proximal promoter regions Yip, K. *et al.* (2012). DH-HMM tends to predict a moderate number of footprints per DHS region, i.e. more footprints than Boyle and fewer footprints than Neph. Interestingly, both Boyle and Neph clearly underestimate the number of footprints in LOT regions, where at least one footprint should be available. This indicates that these methods fail to annotate footprints in LOT regions.

Furthermore, we investigated the DHS regions that contained at least one footprint and the ones that did not overlap with footprints in cell type K562. The Table 13 shows a summary

of these regions. We observed that DH-HMM with three histone modifications missed to annotate a lower number of DHS regions than the competing segmentation-based methods. Moreover, we observed that 1453(97.65%) and 1347(90.52%) regions which the DH-HMM failed to annotate, also did not contain footprints given Boyle and Neph, respectively. To get further insights about these regions, we plotted a histogram of the average DNase-seq read counts inside DHS regions with and without footprints for each method (Fig. 11). The histograms indicate that Boyle and Neph fail to annotate footprints on regions with lower DNase-seq read counts on average. Also, the right tail of Boyle's and Neph's DNase-seq read count distribution on DHS regions without footprint extends slightly further when compared to DH-HMM's. This shows that Boyle and Neph also misses to annotate a few DHS regions with higher DNase-seq read count average.

Finally, we ran the MEME-CHIP tool (Machanick and Bailey, 2011) in the set of DHS regions without any overlap with the footprint predictions from DH-HMM (with three histone modifications) in order to verify if there are motifs enriched in this region. We report, in Table 14, the motif logo, MEME *e*-value (log likelihood ratio approximation given a 0-order Markov background model), percentage of DHS regions that contained such motif (hits) and the putative TFs (determined with TOMTOM tool using Jaspar dataset). We report all motifs with an *e*-value < 0 . At first glance we observed than only three motifs occurred in more than 10% of input regions, out of five enriched motifs found. These motifs present a low number of hits (33.9%, 20.7% and 24.2%, sorted increasingly by *e*-value). The first motif was associated with G rich sequence and eight known TFs. The second motif could not be associated with any known TF. This indicates that the DHS regions without footprints have no strong sequence preference.

4 EXPERIMENTAL DESIGN OF PREVIOUS APPROACHES

4.1 Boyle et al. Method

The predictions (i.e. footprints) based on Boyle, A. P. *et al.* (2011) method were obtained in Furey (2013) for H1-hESC and K562 cell types. They made available the coordinates of the binding locations, which were used exactly as reported. In order to make the comparison with our DH-HMM models fair, we extended their footprints by the same number of base pairs as we extended the footprints generated with the DH-HMM method (5 bp in each direction).

4.2 Neph et al. Method

For the cell type K562, we obtained the digital genomic footprint predictions in Neph, S. *et al.* (2012), which were used exactly as reported. As footprint predictions were not available for the cell type H1-hESC, we obtained the scripts and executed their method with parameters used by the authors in their paper. Briefly, we used the DHS counts as input with the following parameters: the flanking component length varied between 3 and 10, and the central footprint region length varied between 6 and 40. Afterwards, the footprints were filtered by the footprint occupancy score (FOS). As reported in Neph, S. *et al.* (2012), a $FOS \leq 0.95$ generally agrees with a false discovery rate (FDR) of 1%. All other procedures were executed as

described in Neph, S. *et al.* (2012). Additionally, we extended their footprints (both H1-hESC and K562) by the same number of base pairs as we extended the footprints generated with the DH-HMM method (5 bp in each direction).

4.3 Centipede

We obtained the Centipede source code and executed as described in Pique-Regi, R. *et al.* (2011) for all TFs binding in the cell types H1-hESC and K562. Briefly, we fetched the DHS signal surrounding a 200 bp window centered in each motif from our gold standard. The total number of reads within these regions is modeled as a negative binomial distribution and the spatial configuration of reads (i.e. the signal at each genomic position) is modelled as a multinomial (conditional on the total number of reads). Additionally, we obtained PhastCons conservation score (placental mammals on the 46-way multiple alignment) (Siepel, A. *et al.*, 2005) and Ensembl gene annotation from ENCODE (Hubbard, T. *et al.*, 2002) to create the prior probabilities in addition to the MPBSs bit-score. The latter is used to generate a signal that corresponds to the distance of every genomic coordinate to the closest TSS as described in (Pique-Regi, R. *et al.*, 2011). Then, we used the Centipede software to calculate the posterior probabilities of that region being bound by a TF.

We have noticed that Centipede was very sensitive to the level of shrinkage of multinomial and negative binomial parameters on a TF/cell type specific way. Fig. 12 shows examples of AUC variation by varying the level of shrinkage of multinomial and negative binomial parameters from 0.0 to 1.0 by 0.5. We observe high AUC changes, in the same parameter settings, when Centipede is applied to distinct TFs. For instance, when the level of shrinkage of multinomial parameters (L) is set to 0.5 and the level of shrinkage of negative binomial parameters (N) is set to 0.0 we observe optimal results for the TFs CTCF and GABPA in cell type H1-hESC and for TFs CCNT2 and GATA2 in cell type K562. However, for the same parameters, we observe very low AUCs for TFs Myc and SRF in cell type H1-hESC and for TFs NF-E2 and TBP in cell type K562. In another example, when L is set to 1.0 and N is set to 0.0, we observe the best AUC for TFs NF-E2 and TBP but the worst AUC for CCNT2, in cell type K562. Unfortunately, Centipede framework does not provide a procedure for defining such parameters.

Therefore, we decided to perform three evaluation scenarios for Centipede. The first uses the default parameters suggested in Pique-Regi, R. *et al.* (2011), which corresponds to $L = N = 0.0$ (termed 'Default'). The second and third scenarios use L and N parameters estimated in a subset of our gold standard, which includes a random sample of S true MPBSs and S false MPBSs for each TF, where S is the total number of true MPBSs. In both scenarios, 25 AUCs are calculated for each TF corresponding to a grid search consisting on the variation of both L and N parameters from 0.0 to 1.0 by 0.25 intervals. In the second scenario, termed 'Estimated', we use the grid test results (AUCs) to perform a Friedman-Nemenyi hypothesis test for each cell type in order to estimate the best parameters. Then, we use the parameters estimated in the cell type K562 for the predictions made in cell type H1-hESC and vice versa. In the third scenario, termed 'Optimistic', we simply use, for each TF, the parameter settings that generated the best AUC. It is important to point out that this last TF and cell specific selection criterion is not applicable on a real scenario, where ChIP-seq experiments for the

tested TFs are not available. This was performed only to obtain the method accuracies' upper boundaries for comparison purposes.

The Fig. 13 shows the AUC distribution based on all three scenarios discussed previously. In both cell types, the 'Default' model performed poorly than the other two. Also, we observe that the 'Estimated' model AUC distribution is close, but still significantly lower than the 'Optimistic' model AUC distribution (p -value $\leq 10^{-5}$ with Wilcoxon-Mann-Whitney test). Interestingly, the set of parameters estimated in the cell type specific procedure were quite similar ($L = 0.75$ and $N = 0$ for H1-hESC and $L = 0.75$ and $N = 0.25$ for K562). Our interpretation is that given our large evaluation set (about 5 times more TFs than in Pique-Regi, R. *et al.* (2011)), these parameter choices are more robust and should be used as default parameters to Centipede.

4.4 Cuellar-Partida et al. Method

We obtained Cuellar-Partida, G. *et al.* (2012) scripts to generate priors and executed as described in their paper. To create DHS input data, we evaluated the number of reads aligning to a window of 150 bp, specified every 20 bp. Histone modification input data was specified in a 25 bp resolution. In every position, it was summed 1 if a mapped read fell within 0–200 bp from the 25 bp window and 0.25 if it occurred within 200–300 bp. In order to compare their strategy to ours, priors were created using, in combination, DHS and histone modifications H3K4me1+H3K4me3 and H3K4me1+H3K4me3+H3K9a, which we refer as Cuellar(2) and Cuellar(3).

To obtain the predictions, the program FIMO (Grant, C. E. *et al.*, 2011) was used along with the priors, as suggested by the authors. The establishment of the motif matching p -value threshold led to an issue similar to the selection of Centipede parameters (Section 4.3), i.e. it was used a selection criteria based on a *a posteriori* evaluation of the AUCs obtained when using different FIMO p -value thresholds (10^{-3} , 10^{-4} , 10^{-5} , 10^{-6} and 10^{-7}).

In Fig. 14 we can observe that the distribution of AUCs when FIMO p -value = 10^{-5} is used seems to outperforms all other thresholds. In order to test this claim, we performed a Friedman-Nemenyi test on the results of each FIMO p -value threshold. By observing the Friedman ranking and the Friedman-Nemenyi results available in Tables 15 and 16, respectively, we are able to define the FIMO p -value = 10^{-5} as the best choice.

REFERENCES

- Bergman, C. M. *et al.* (2005). Drosophila DNase i footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, drosophila melanogaster. *Bioinformatics*, **21**(8), 1747–1749.
- 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.
- Crawford, G. E. *et al.* (2006). Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). *Genome Research*, **16**(1), 123–131.
- Cuellar-Partida, G. *et al.* (2012). Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, **28**(1), 56–62.
- de Wit, E. *et al.* (2013). The pluripotent genome in three dimensions is shaped around pluripotency factors. *Nature*, **501**(7466), 227–231.
- ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.
- Furey, T. S. (2013). Furey lab: Dnase-seq footprints. <http://fureylab.web.unc.edu/datasets/footprints/>.

- Grant, C. E. *et al.* (2011). FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**(7), 1017–1018.
- Hesselberth, J. R. *et al.* (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature Methods*, **6**(4), 283–289.
- 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*.
- Hubbard, T. *et al.* (2002). The ensembl genome database project. *Nucleic acids research*, **30**(1), 38–41.
- Jankowski, A. *et al.* (2013). Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers. *Genome Research*.
- Kulakovskiy, I. V. *et al.* (2009). Motif discovery and motif finding from genome-mapped DNase footprint data. *Bioinformatics*, **25**(18), 2318–2325.
- Kundaje, A. *et al.* (2012). Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome Research*, **22**(9), 1735–1747.
- Machanick, P. and Bailey, T. L. (2011). MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics*, **27**(12), 1696–1697.
- Mathelier, A. *et al.* (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, **42**(D1), D142–D147.
- Matys, V. *et al.* (2006). TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, **34**(Database issue), D108–D110.
- 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.
- Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, **77**(2), 257–286.
- Rada-Iglesias, A. *et al.* (2010). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature*, **470**, 279–283+.
- Robasky, K. and Bulyk, M. L. (2011). UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic acids research*, **39**(Database issue).
- Siepel, A. *et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, **15**(8), 1034–1050.
- Song, L. and Crawford, G. E. (2010). DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. *Cold Spring Harbor Protocols*, **2010**(2), pdb.prot5384+.
- 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.
- Whittington, T. *et al.* (2009). High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Research*, **37**(1), 14–25.
- Wilczynski, B. *et al.* (2009). Finding evolutionarily conserved cis-regulatory modules with a universal set of motifs. *BMC bioinformatics*, **10**(1), 82+.
- Wingender, E. *et al.* (2013). TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Research*, **41**(D1), D165–D170.
- Won, K. J. *et al.* (2010). Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biology*, **11**(1), R7+.
- 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+.

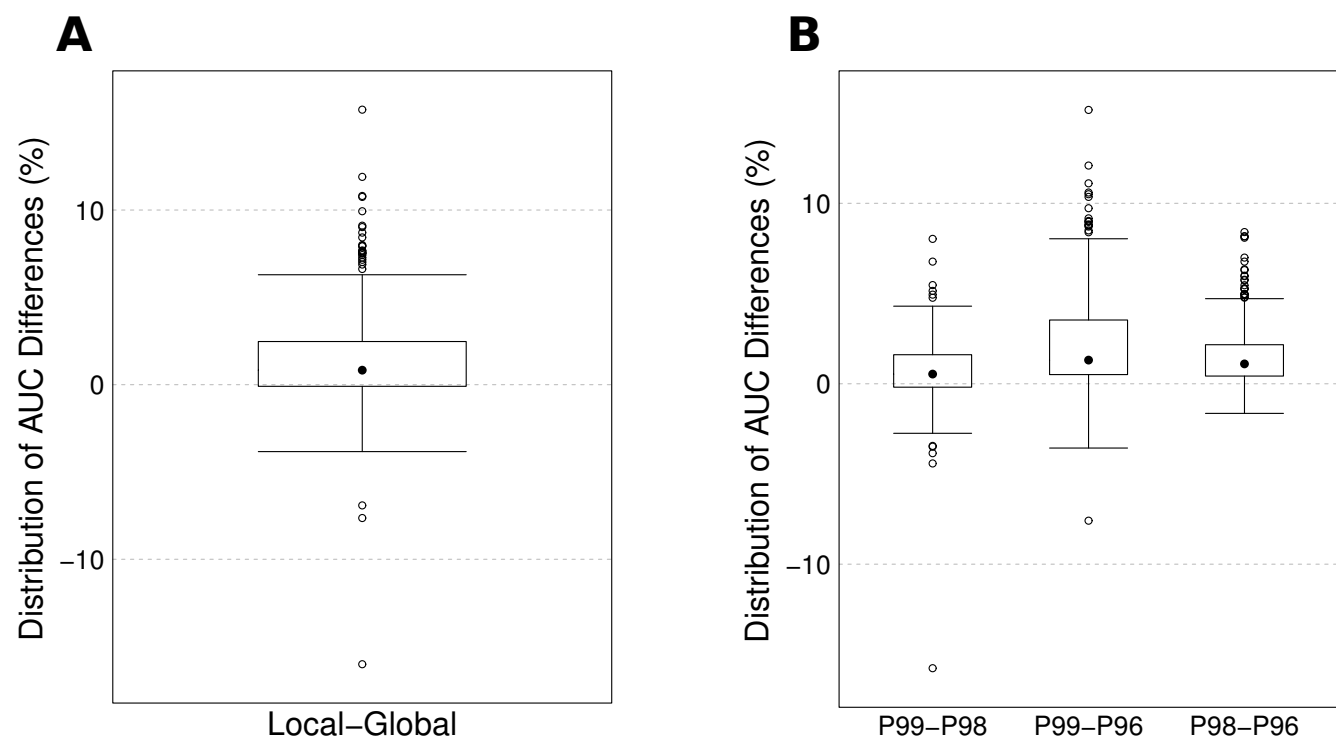


Fig. 1. Scaling parameter grid test. **(A)** Distribution of the difference between local and global AUCs for all percentile values tested. Positive values indicate the advantage of the local approach. **(B)** Distribution of the difference between specific percentile values given the local approach. Positive values indicate the advantage of higher percentiles.

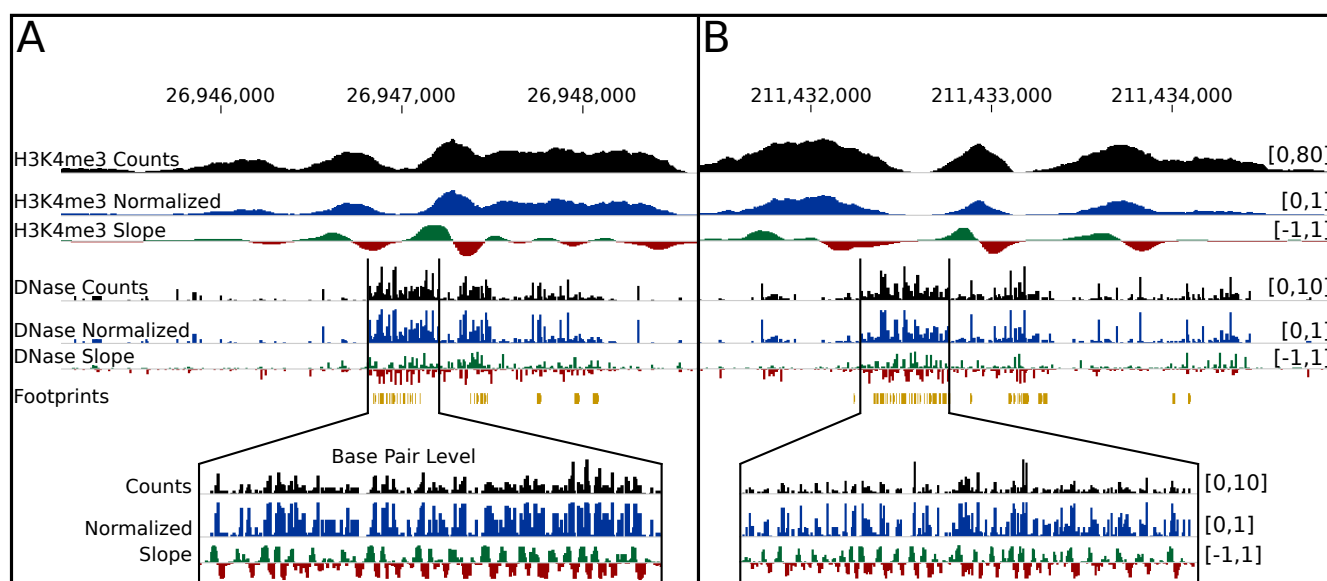


Fig. 2. Example of count, normalized and slope signals for H3K4me3 and DHS for two distinct regions using data from H1-hESC cell type. Signals' ranges can be seen in the right part of the figure.

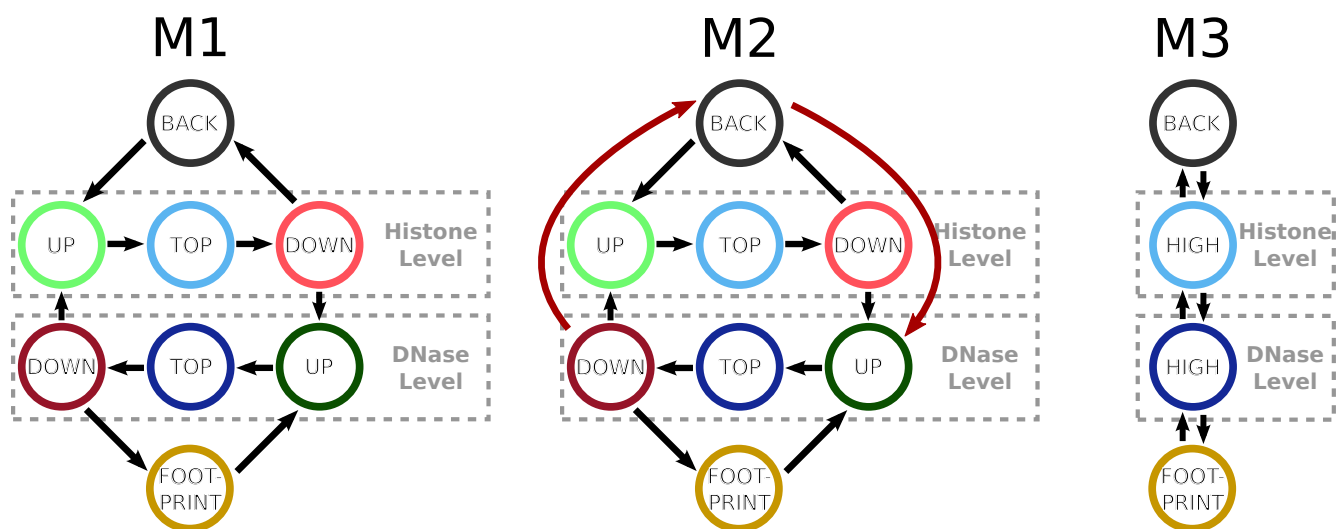


Fig. 3. Depiction of all HMM topologies tested.

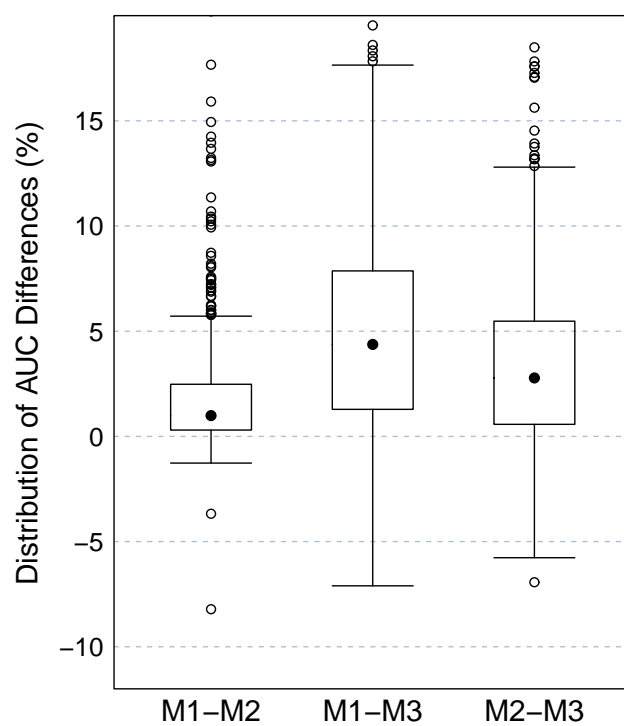


Fig. 4. Distribution of AUC differences between all HMM topologies tested (M1, M2 and M3).

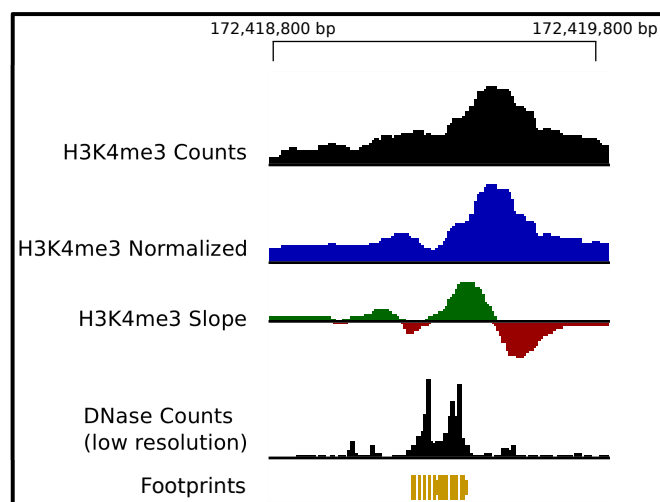


Fig. 5. Example of H3K4me3 count, normalized and slope signals, DHS count signal and footprints predicted in cell type K562 in a region with asymmetrical histone modification profile. Although the leftmost peak from the peak-dip-peak pattern of the H3K4me3 contains very small count signals (very close to the background signal, in this region), the DH-HMM model M1 is still capable of predicting the footprints within the DHS region. Note that after the normalization of the H3K4me3 counts, the resulting signal delineates the DHS region more clearly. We draw the attention to the fact that the DHS signal zooming is such that we are not able to visually identify the footprint signatures.

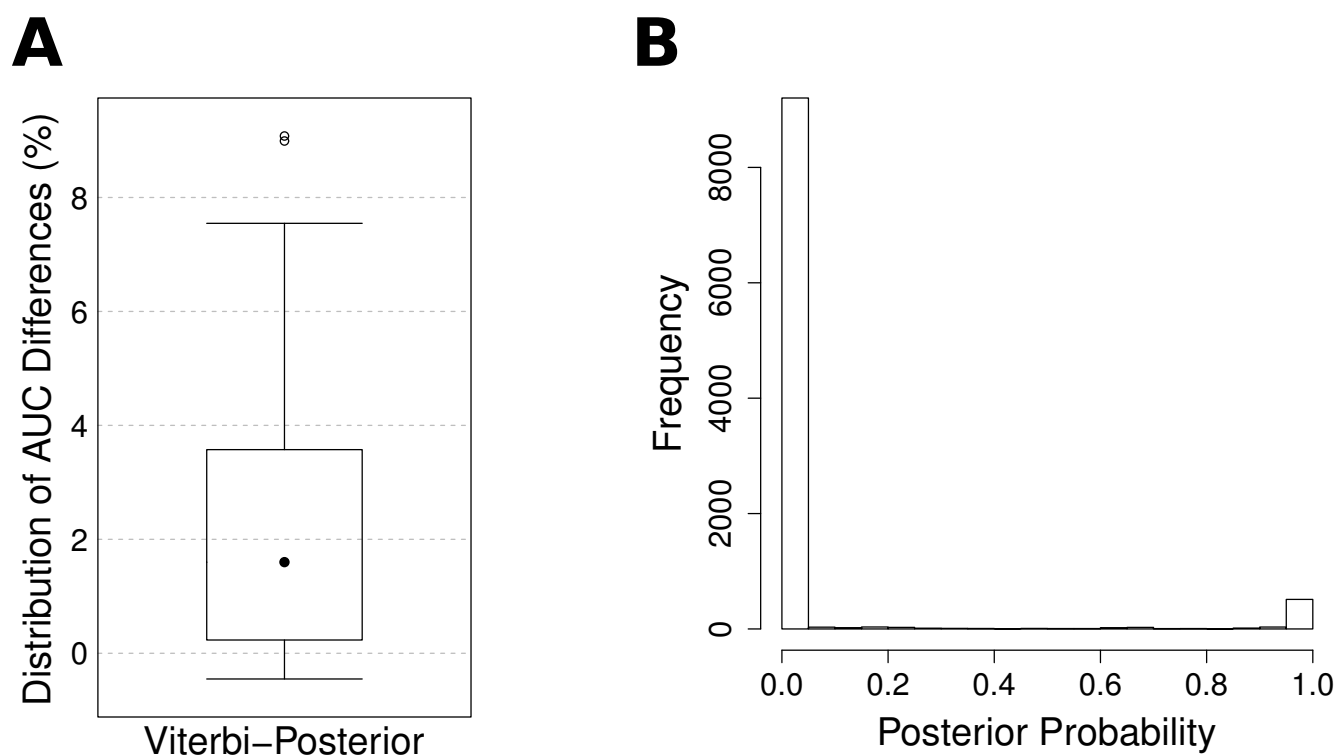


Fig. 6. Viterbi vs. posterior probability test. **(A)** Distribution of AUC differences between the predictions made using the Viterbi algorithm and the posterior probability. **(B)** Distribution of the posterior probability of the HMM being in the state `FOOTPRINT` evaluated in a DHS region in chromosome 1 from 211,431,582 to 211,434,492 bp.

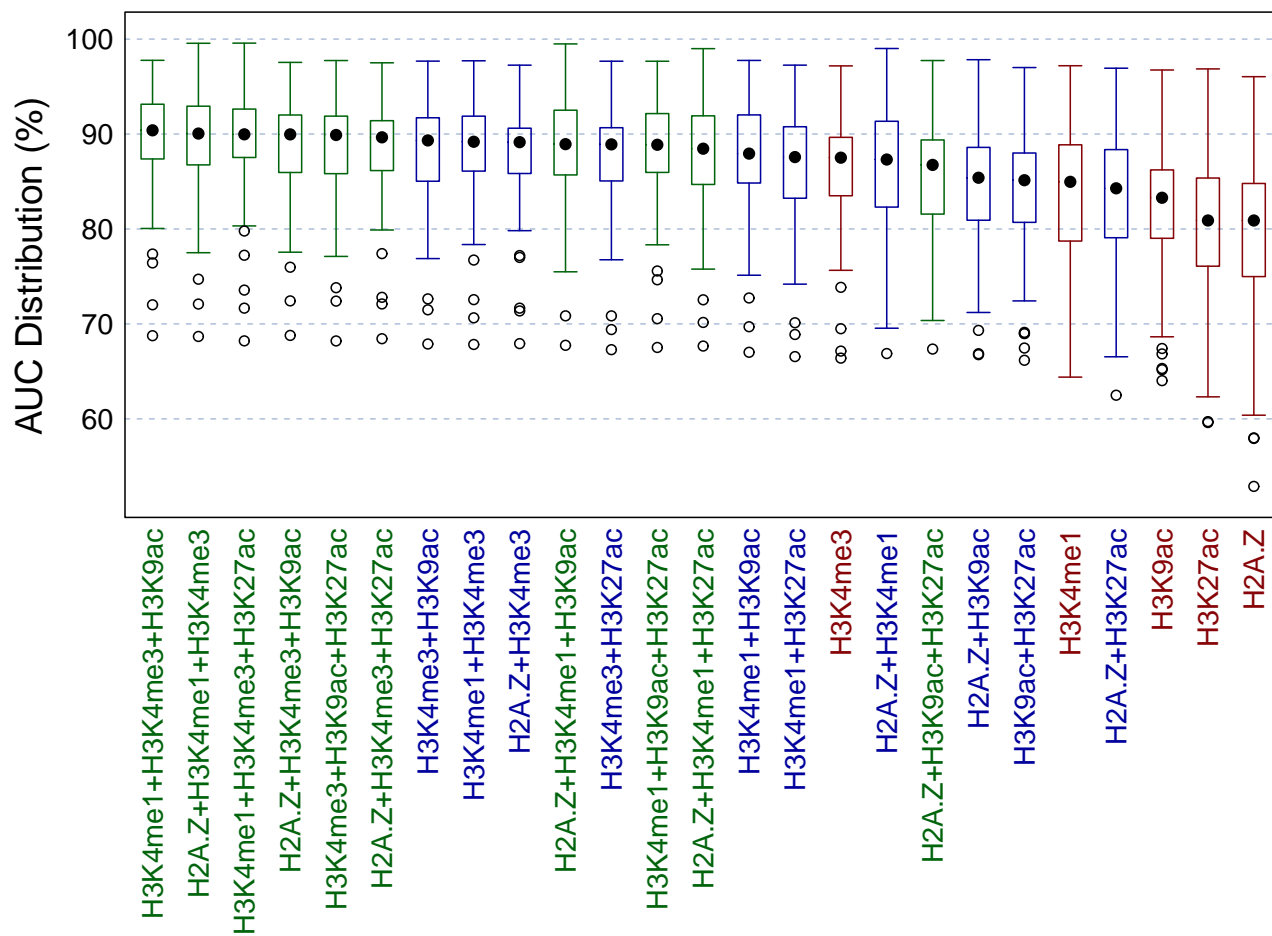


Fig. 7. Distribution of AUCs for all histone modification models tested. Boxplots are sorted in decreasing order according to their median. Models representing histone triples, pairs and singles are colored in green, blue and red, respectively.

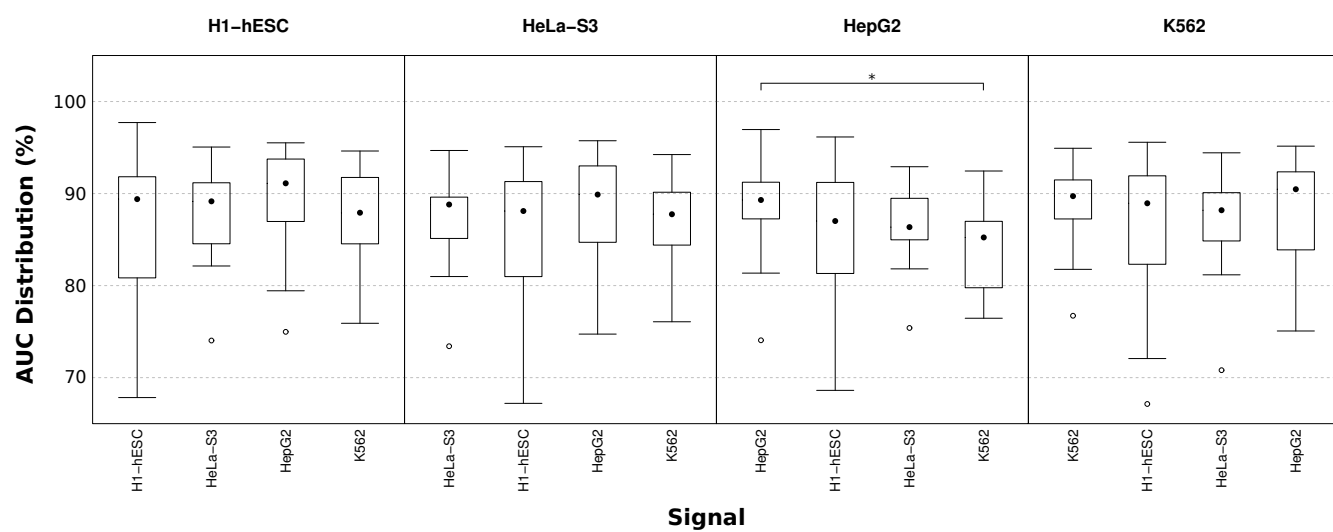


Fig. 8. Distribution of AUCs for all HMM models (top x-axis labels) and applied to all cell types (bottom x-axis labels). The first boxplot within each set represents the model trained in the same cell type as the one it was applied to. Statistical significance on the pairwise difference between these distributions is represented by the three-star system.

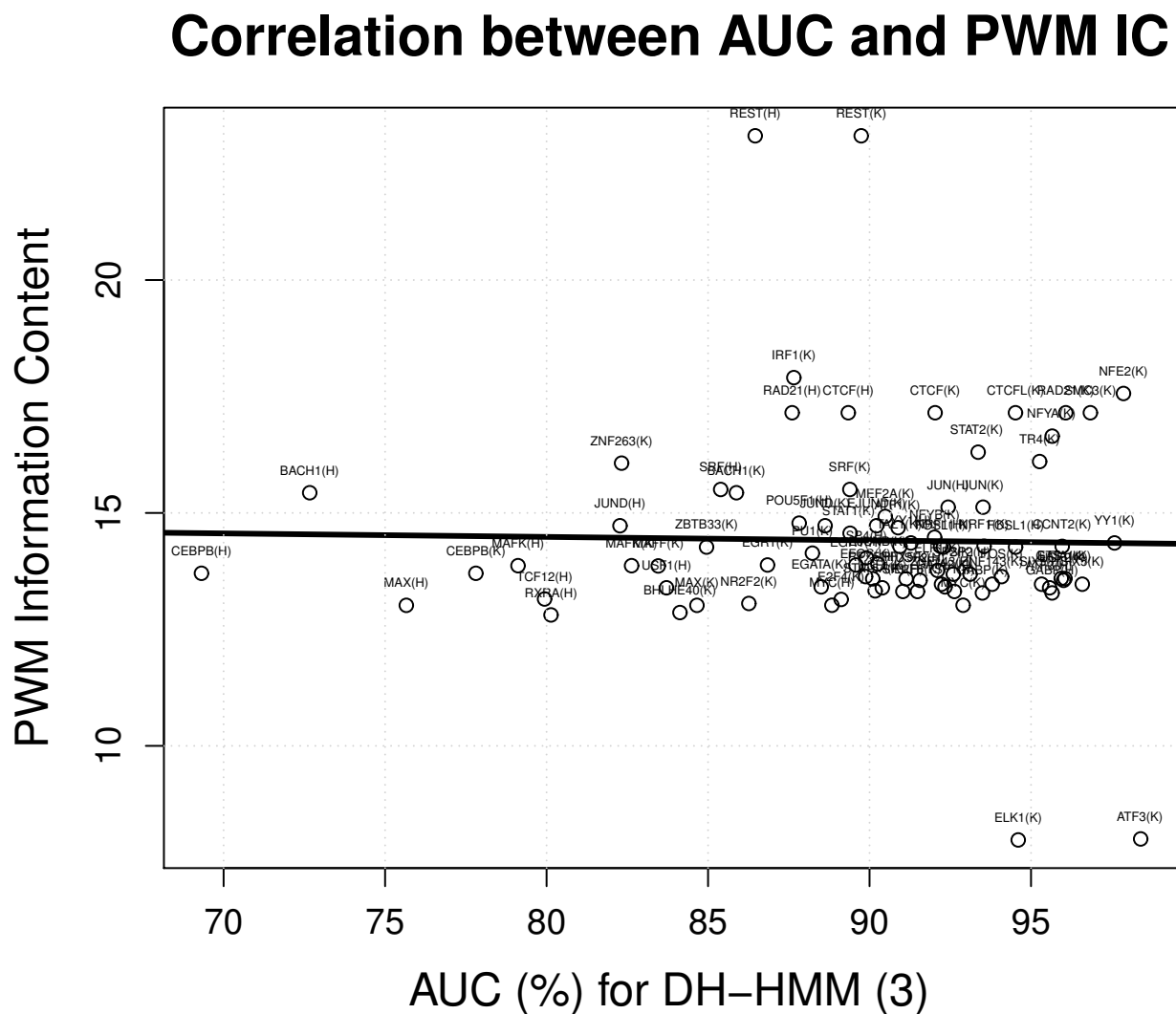


Fig. 9. Correlation between our method's AUC for DH-HMM(3) and the PWM's IC. Labels with each TF's name are shown with '(H)' for factors binding in H1-hESC and '(K)' for factors binding in K562. The HMM models were trained in the same cell type they were applied. The thicker line in the scatterplot represents the regression line.

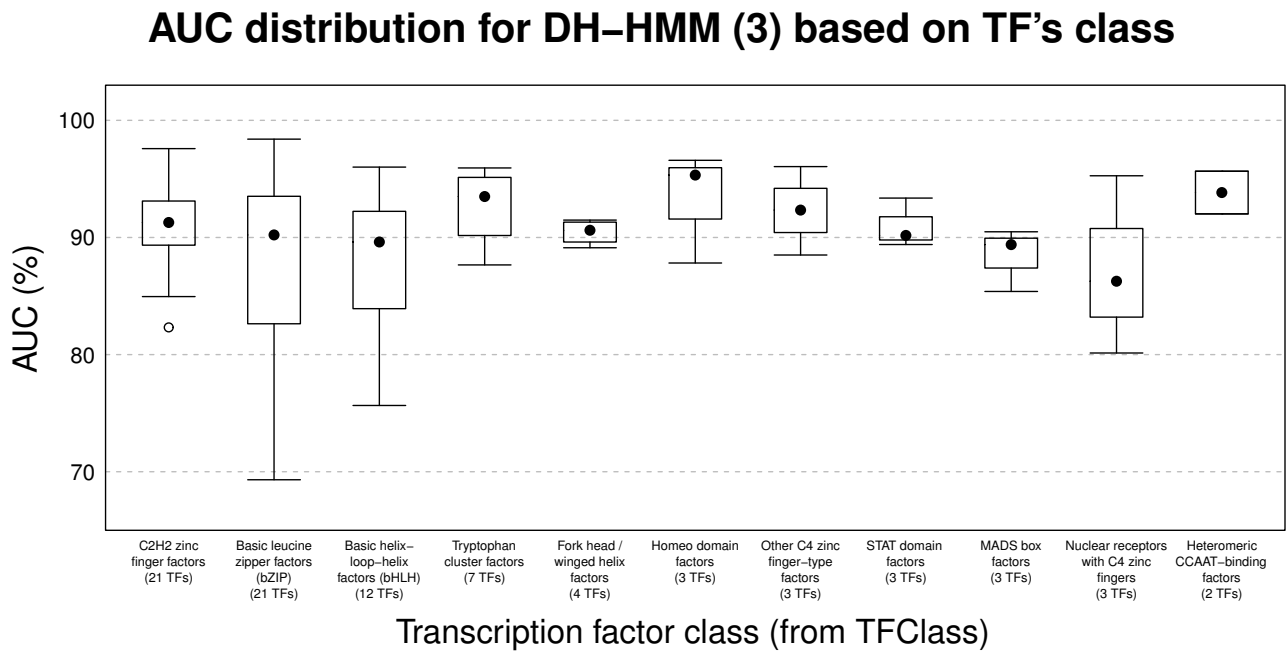


Fig. 10. Distribution of AUCs for our method (with three histone modifications) categorized by TF class (using TFClass). Below each category's label in the x-axis we show the number of TFs that belong to that class. The boxplot is sorted decreasingly by the number of the TFs in each class.

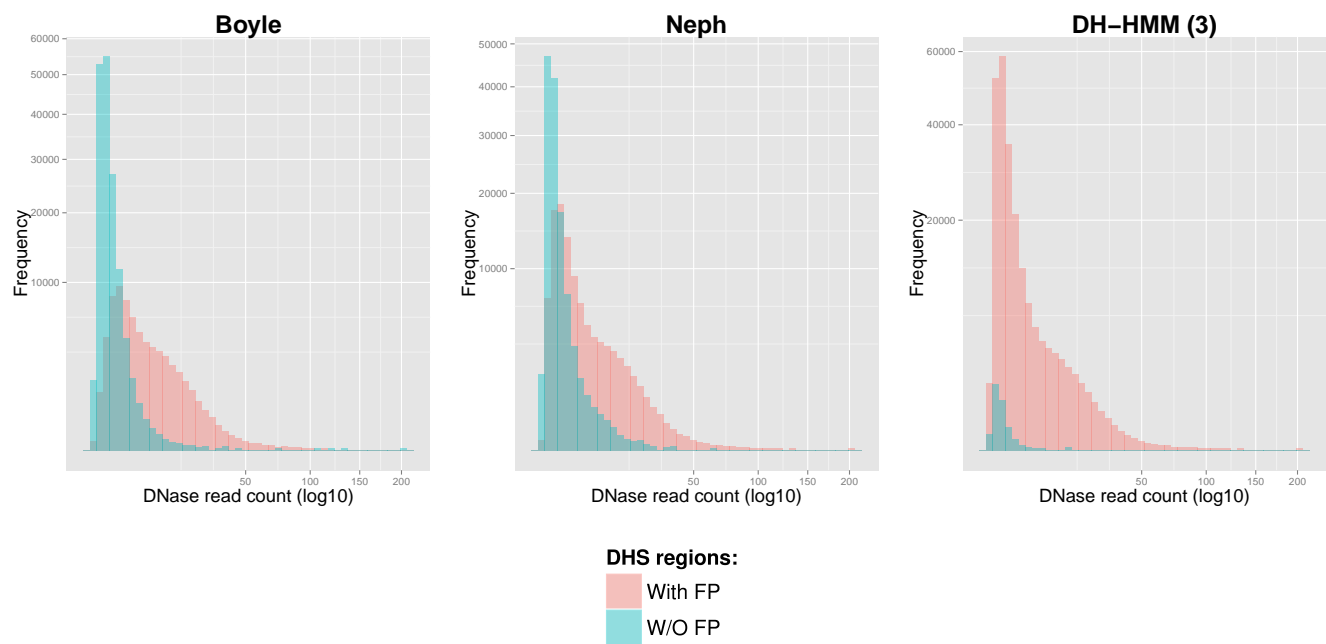


Fig. 11. Distribution of the average DNase-seq read counts (in \log_{10} scale) inside DHS regions with/without footprints for the main segmentation-based methods.

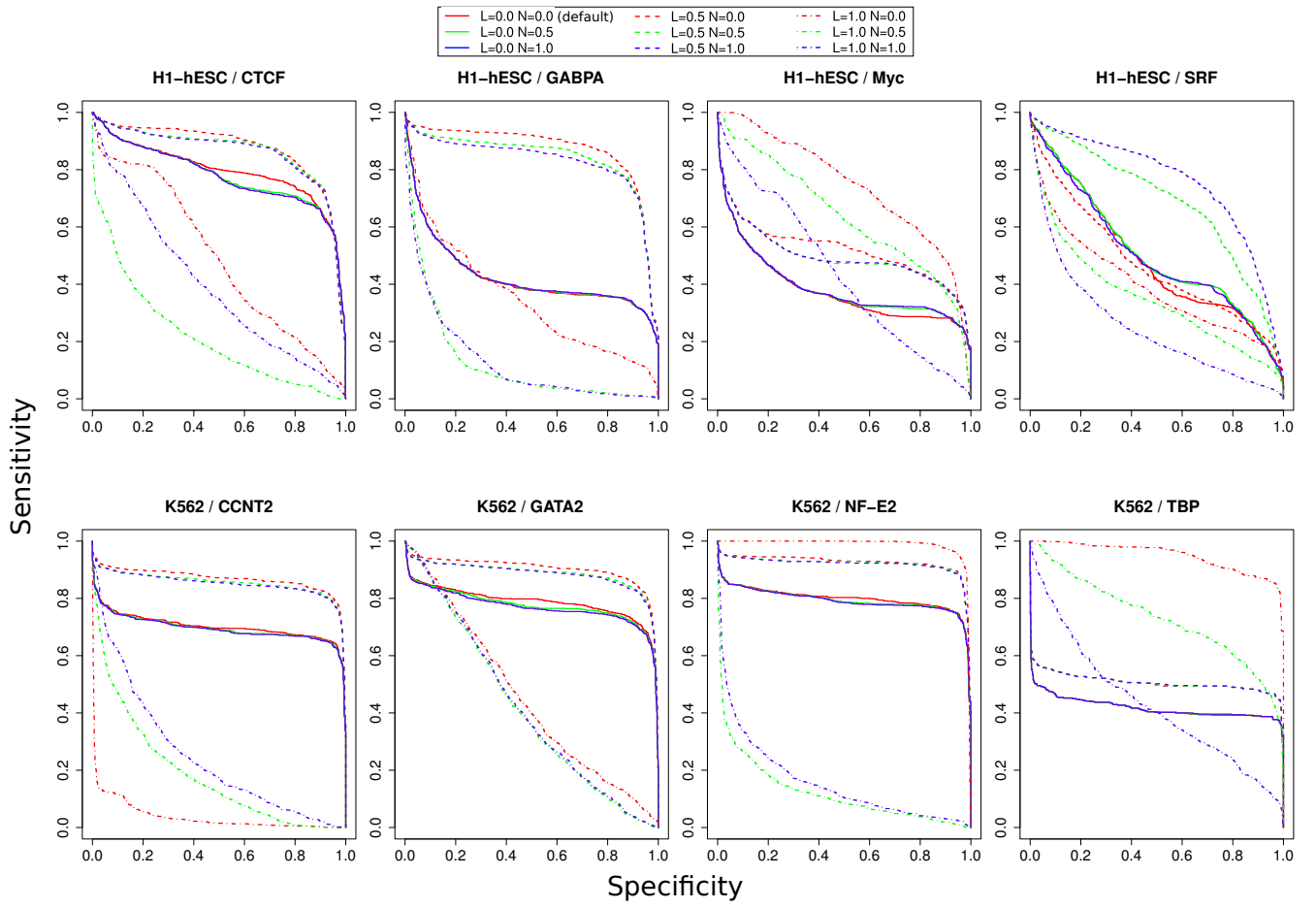


Fig. 12. Centipede's ROC-like curves created for multiple TFs based on a grid variation of Centipede's level of shrinkage of multinomial parameters (L) and level of shrinkage of negative binomial parameters (N) from 0.0 to 1.0 by 0.5.

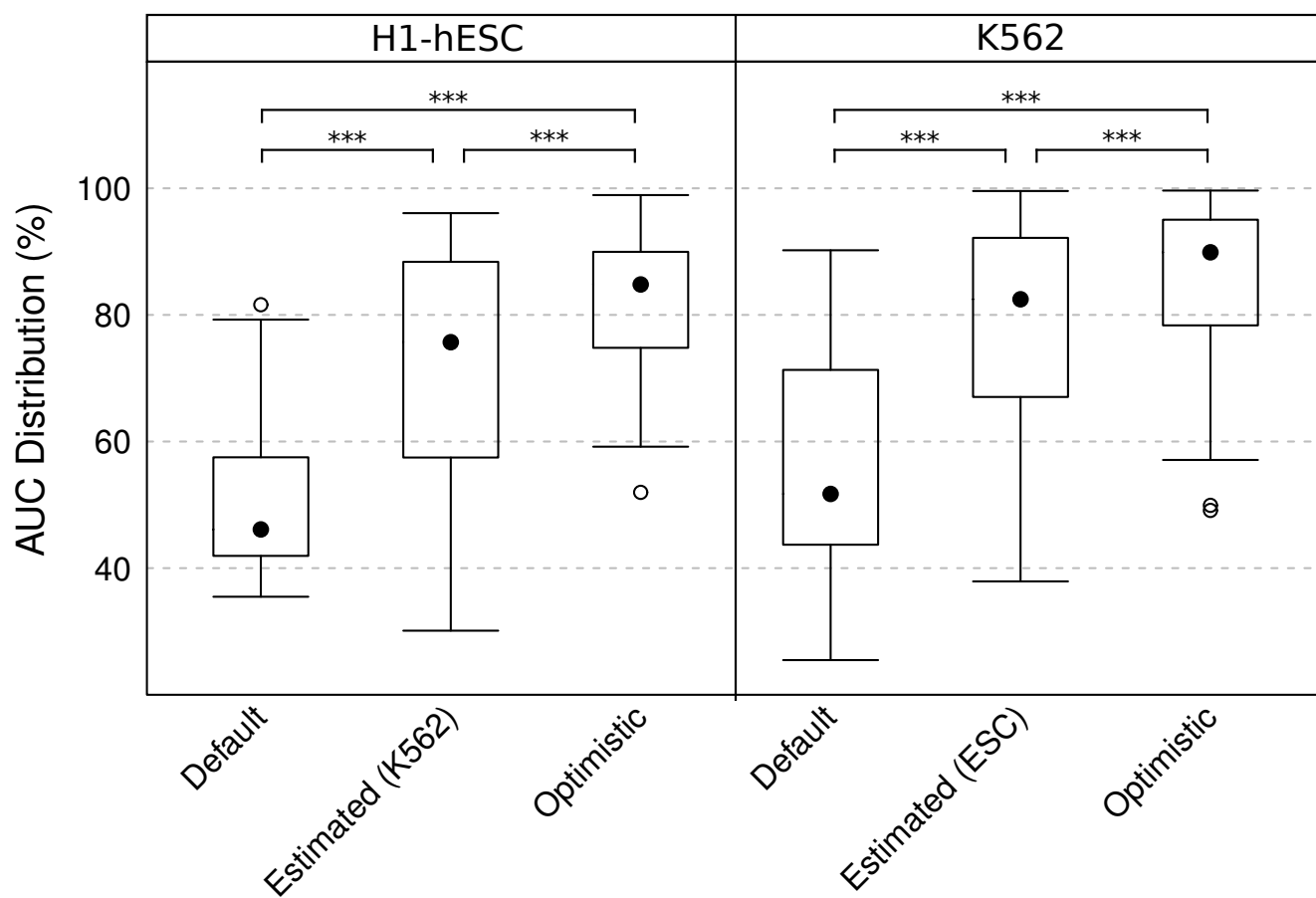


Fig. 13. Distribution of AUCs for the three scenarios in which Centipede's level of shrinkage of multinomial parameters (L) and level of shrinkage of negative binomial parameters (N) were defined. 'Default' refers to $L = N = 0.0$. 'Estimated' refers to parameters estimated on a different cell type. 'Optimistic' refers to parameters estimated on the same cell type for each TF individually. Statistical significance on the pairwise difference between these distributions is represented by the three-star system.

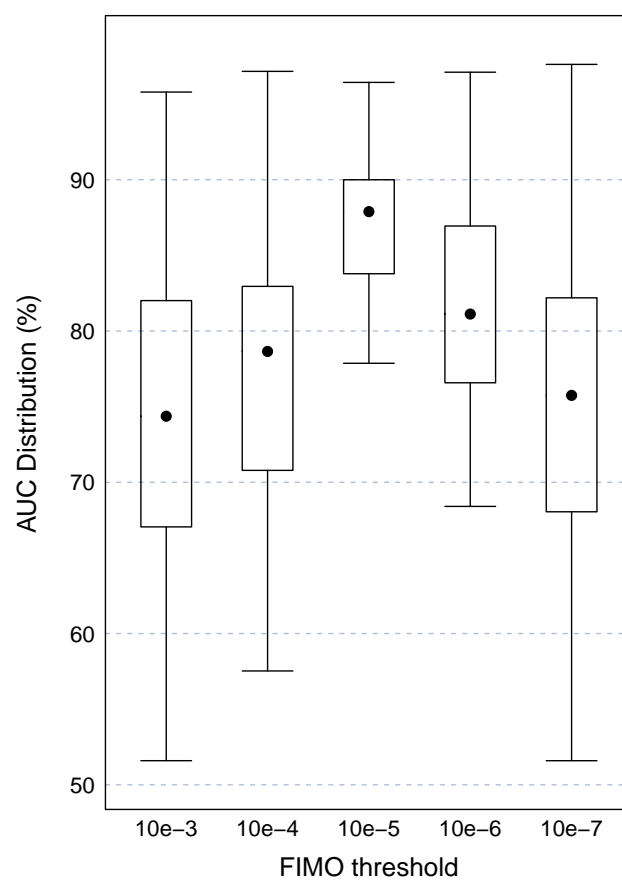


Fig. 14. Distribution of AUCs when using Cuellar's priors with different FIMO p -value thresholds.

Table 1. Friedman ranking. For each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

Sensitivity		Specificity		AUC	
Cuellar (2)	2.2213	Boyle	1.0902	DH-HMM (3)	1.8293
Cuellar (3)	2.7459	Neph	2.0082	DH-HMM (2)	2.6707
DH-HMM (3)	3.5	DH-HMM (2)	3.5246	Centipede	4.3902
Centipede	3.5328	Centipede	3.7541	Neph	4.8537
DH-HMM (2)	4.8852	DH-HMM (3)	4.6557	Cuellar (2)	5.1098
H-HMM (3)	4.918	Cuellar (2)	6.9508	Cuellar (3)	5.3293
H-HMM (2)	6.7705	H-HMM (2)	7.1475	H-HMM (3)	6.5732
Neph	7.541	Cuellar (3)	7.2623	H-HMM (2)	7.3171
Boyle	8.8852	H-HMM (3)	8.6066	Boyle	7.5

Table 2. Transition probabilities of the HMM trained with H3K4me3 using H1-hESC data. Transitions are specified from the states in the rows to the states in the columns. Histone level states are denoted with '(H)' and DNase level states with '(D)'. The FOOTPRINT state is abbreviated as 'FP'.

	BACK	UP (H)	TOP (H)	DOWN (H)	UP (D)	TOP (D)	DOWN (D)	FP
BACK	0.9997	0.0003	0.0	0.0	0.0	0.0	0.0	0.0
UP (H)	0.0	0.9915	0.0085	0.0	0.0	0.0	0.0	0.0
TOP (H)	0.0	0.0	0.9901	0.0099	0.0	0.0	0.0	0.0
DOWN (H)	0.0057	0.0	0.0	0.9861	0.0082	0.0	0.0	0.0
UP (D)	0.0	0.0	0.0	0.0	0.6515	0.3485	0.0	0.0
TOP (D)	0.0	0.0	0.0	0.0	0.0	0.783	0.217	0.0
DOWN (D)	0.0	0.0339	0.0	0.0	0.0	0.0	0.577	0.3891
FP	0.0	0.0	0.0	0.0	0.0564	0.0	0.0	0.9436

Table 3. Signals' mean values for each state of the HMM trained with H3K4me3 using H1-hESC data. Histone level states are denoted with '(H)' and DNase level states with '(D)'. The FOOTPRINT state is abbreviated as 'FP'.

	DNase norm.	DNase slope	Histone norm.	Histone slope
BACK	0.0045	-0.0002	0.0441	0.0007
UP (H)	0.0501	0.0043	0.1983	0.2995
TOP (H)	0.0445	-0.0075	0.4693	0.0158
DOWN (H)	0.0636	0.0003	0.2309	-0.4237
UP (D)	0.1537	0.6343	0.0894	-0.0647
TOP (D)	0.4244	0.0059	0.1091	-0.0735
DOWN (D)	0.1578	-0.6562	0.0816	-0.0434
FP	0.0902	-0.0162	0.1009	-0.0436

Table 4. Covariance matrices for each state of the HMM trained with H3K4me3 using H1-hESC data. Within each state's matrix, lines and rows are sorted by signal type as DNase normalized, DNase slope, H3K4me3 normalized and H3K4me3 slope. Histone level states are denoted with '(H)' and DNase level states with '(D)'. The FOOTPRINT state is abbreviated as 'FP'.

BACK	0.0025	-0.0001	0.0001	0.0	UP (H)	0.0222	0.0001	0.003	0.0057
	-0.0001	0.0025	0.0	0.0		0.0001	0.0155	0.0006	0.0005
	0.0001	0.0	0.0047	0.0		0.003	0.0006	0.0101	0.0105
	0.0	0.0	0.0	0.0019		0.0057	0.0005	0.0105	0.0341
TOP (H)	0.0216	0.0003	-0.0009	0.0014	DOWN (H)	0.0239	0.0001	-0.0033	-0.0002
	0.0003	0.0196	0.0005	0.0003		0.0001	0.009	0.0002	-0.0006
	-0.0009	0.0005	0.0047	-0.001		-0.0033	0.0002	0.0156	-0.0095
	0.0014	0.0003	-0.001	0.0193		-0.0002	-0.0006	-0.0095	0.0313
UP (D)	0.0705	0.0246	-0.0053	0.0025	TOP (D)	0.1559	-0.002	-0.0079	0.0052
	0.0246	0.0714	-0.0038	-0.0015		-0.002	0.0384	-0.0008	0.0021
	-0.0053	-0.0038	0.0045	-0.0056		-0.0079	-0.0008	0.007	-0.0096
	0.0025	-0.0015	-0.0056	0.0125		0.0052	0.0021	-0.0096	0.0184
DOWN (D)	0.0687	-0.011	-0.0048	0.004	FP	0.0358	-0.0019	-0.0025	0.0007
	-0.011	0.055	0.0039	-0.0		-0.0019	0.0225	0.0001	0.0002
	-0.0048	0.0039	0.0039	-0.0044		-0.0025	0.0001	0.0068	-0.0069
	0.004	-0.0	-0.0044	0.0109		0.0007	0.0002	-0.0069	0.0121

Table 5. Coverage of DHS, H3K4me1 and H3K4me3 enriched regions for H1-hESC and K562 cell types.

Data Type	H1-hESC	K562
DHS	87294396 (2.65%)	53487366 (1.62%)
H3K4me1	16782633 (0.51%)	150487948 (4.56%)
H3K4me3	30857168 (0.94%)	71126062 (2.16%)
H3K4me1+ H3K4me3	45289399 (1.37%)	174412067 (5.29%)
DHS+ H3K4me1+ H3K4me3	105941410 (3.21%)	191774248 (5.81%)

Table 6. Friedman ranking regarding AUC for models based on different histone modification combinations. The models are displayed in decreasing order with their respective Friedman ranking. **In this analysis, it was used data from the combination of cell types H1-hESC and K562.**

AUC Friedman Ranking	
H3K4me1+H3K4me3+H3K9ac	3.5568
H3K4me1+H3K4me3+H3K27ac	4.8182
H2A.Z+H3K4me1+H3K4me3	5.0341
H2A.Z+H3K4me3+H3K9ac	6.4318
H3K4me3+H3K9ac+H3K27ac	6.9432
H2A.Z+H3K4me3+H3K27ac	7.7841
H3K4me1+H3K4me3	8.625
H3K4me1+H3K9ac+H3K27ac	8.8068
H2A.Z+H3K4me1+H3K9ac	9.2614
H3K4me3+H3K9ac	9.4318
H2A.Z+H3K4me1+H3K27ac	10.8409
H3K4me3+H3K27ac	11.0795
H2A.Z+H3K4me3	11.5227
H3K4me1+H3K9ac	12.0227
H3K4me1+H3K27ac	14.3636
H2A.Z+H3K4me1	15.2386
H2A.Z+H3K9ac+H3K27ac	15.6705
H3K4me3	16.1932
H2A.Z+H3K9ac	18.8864
H3K9ac+H3K27ac	18.9545
H3K4me1	19.9432
H2A.Z+H3K27ac	20.1364
H3K9ac	22.1818
H3K27ac	23.5568
H2A.Z	23.7159

Table 7. Friedman-Nemenyi hypothesis test results for the AUC metric and the models based on different histone modifications. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1. **In this analysis, it was used data from the combination of cell types H1-hESC and K562.**

	H3K4me1+H3K4me3+H3K9ac	H3K4me1+H3K4me3+H3K27ac	H2A.Z+H3K4me1+H3K4me3	H2A.Z+H3K4me3+H3K9ac	H3K4me3+H3K9ac+H3K27ac	H2A.Z+H3K4me3+H3K27ac	H3K4me1+H3K4me3	H3K4me1+H3K9ac+H3K27ac	H2A.Z+H3K4me1+H3K9ac	H3K4me3+H3K9ac	H2A.Z+H3K4me1+H3K27ac	H3K4me3+H3K27ac	H2A.Z+H3K4me3	H3K4me1+H3K9ac	H3K4me1+H3K27ac	H2A.Z+H3K4me1	H2A.Z+H3K9ac+H3K27ac	H3K4me3	H2A.Z+H3K9ac	H3K9ac+H3K27ac	H3K4me1	H2A.Z+H3K27ac	H3K9ac	H3K27ac	H2A.Z
H3K4me1+H3K4me3+H3K9ac																									
H3K4me1+H3K4me3+H3K27ac																									
H2A.Z+H3K4me1+H3K4me3																									
H2A.Z+H3K4me3+H3K9ac																									
H3K4me3+H3K9ac+H3K27ac																									
H2A.Z+H3K4me3+H3K27ac																									
H3K4me1+H3K4me3																									
H3K4me1+H3K9ac+H3K27ac																									
H2A.Z+H3K4me1+H3K9ac																									
H3K4me3+H3K9ac																									
H2A.Z+H3K4me1+H3K27ac																									
H3K4me3+H3K27ac																									
H2A.Z+H3K4me3																									
H3K4me1+H3K9ac																									
H3K4me1+H3K27ac																									
H2A.Z+H3K4me1																									
H2A.Z+H3K9ac+H3K27ac																									
H3K4me3																									
H2A.Z+H3K9ac																									
H3K9ac+H3K27ac																									
H3K4me1																									
H2A.Z+H3K27ac																									
H3K9ac																									
H3K27ac																									
H2A.Z																									

Table 8. Friedman ranking regarding AUC for models based on different histone modification combinations. The models are displayed in decreasing order with their respective Friedman ranking. In this analysis, it was used data from the cell types H1-hESC and K562 in separate.

AUC (H1-hESC)		AUC (K562)	
H2A.Z+H3K4me1+H3K4me3	3.5172	H3K4me1+H3K4me3+H3K9ac	2.7797
H3K4me1+H3K4me3+H3K9ac	5.1379	H3K4me1+H3K4me3+H3K27ac	4.5593
H3K4me1+H3K4me3+H3K27ac	5.3448	H3K4me3+H3K9ac+H3K27ac	5.2542
H2A.Z+H3K4me1+H3K9ac	6.6897	H2A.Z+H3K4me1+H3K4me3	5.7797
H2A.Z+H3K4me1+H3K27ac	6.7931	H2A.Z+H3K4me3+H3K9ac	5.8136
H2A.Z+H3K4me3+H3K27ac	7.5172	H3K4me3+H3K9ac	7.2034
H2A.Z+H3K4me3+H3K9ac	7.6897	H2A.Z+H3K4me3+H3K27ac	7.9153
H3K4me1+H3K9ac+H3K27ac	8.9655	H3K4me1+H3K4me3	8.4237
H3K4me1+H3K4me3	9.0345	H3K4me1+H3K9ac+H3K27ac	8.7288
H3K4me3+H3K9ac+H3K27ac	10.3793	H3K4me3+H3K27ac	9.6271
H2A.Z+H3K4me1	10.7241	H2A.Z+H3K4me1+H3K9ac	10.5254
H2A.Z+H3K4me3	10.931	H2A.Z+H3K4me3	11.8136
H3K4me1+H3K9ac	12.3448	H3K4me1+H3K9ac	11.8644
H3K4me1+H3K27ac	13.3448	H2A.Z+H3K4me1+H3K27ac	12.8305
H2A.Z+H3K9ac+H3K27ac	13.8966	H3K4me3	14.4746
H3K4me3+H3K9ac	13.9655	H3K4me1+H3K27ac	14.8644
H3K4me3+H3K27ac	14.0345	H2A.Z+H3K9ac+H3K27ac	16.5424
H2A.Z+H3K27ac	17.2069	H2A.Z+H3K4me1	17.4576
H2A.Z+H3K9ac	17.4483	H3K9ac+H3K27ac	17.8814
H3K4me1	19.4138	H2A.Z+H3K9ac	19.5932
H3K4me3	19.6897	H3K4me1	20.2034
H3K9ac+H3K27ac	21.1379	H3K9ac	21.3559
H2A.Z	21.7586	H2A.Z+H3K27ac	21.5763
H3K9ac	23.8621	H3K27ac	23.2542
H3K27ac	24.1724	H2A.Z	24.678

Table 9. Friedman-Nemenyi hypothesis test results for the AUC metric and the models based on different histone modifications. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1. In this analysis, it was used data from the cell type **H1-hESC**.

	H2A.Z+H3K4me1+H3K4me3	H3K4me1+H3K4me3+H3K9ac	H3K4me1+H3K4me3+H3K27ac	H2A.Z+H3K4me1+H3K9ac	H2A.Z+H3K4me1+H3K27ac	H2A.Z+H3K4me3+H3K27ac	H2A.Z+H3K4me3+H3K9ac	H3K4me1+H3K9ac+H3K27ac	H3K4me1+H3K4me3	H3K4me3+H3K9ac+H3K27ac	H2A.Z+H3K4me1	H2A.Z+H3K4me3	H3K4me1+H3K9ac	H3K4me1+H3K27ac	H2A.Z+H3K9ac+H3K27ac	H3K4me3+H3K9ac	H3K4me3+H3K27ac	H2A.Z+H3K27ac	H2A.Z+H3K9ac	H3K4me1	H3K4me3	H3K9ac+H3K27ac	H2A.Z	H3K9ac	H3K27ac
H2A.Z+H3K4me1+H3K4me3																									
H3K4me1+H3K4me3+H3K9ac																									
H3K4me1+H3K4me3+H3K27ac																									
H2A.Z+H3K4me1+H3K9ac																									
H2A.Z+H3K4me1+H3K27ac																									
H2A.Z+H3K4me3+H3K27ac																									
H2A.Z+H3K4me3+H3K9ac																									
H3K4me1+H3K9ac+H3K27ac																									
H3K4me1+H3K4me3																									
H3K4me3+H3K9ac+H3K27ac	+																								
H2A.Z+H3K4me1	*																								
H2A.Z+H3K4me3	*																								
H3K4me1+H3K9ac	*	*	+																						
H3K4me1+H3K27ac	*	*	*	+																					
H2A.Z+H3K9ac+H3K27ac	*	*	*	*	*	*																			
H3K4me3+H3K9ac	*	*	*	*	*	*																			
H3K4me3+H3K27ac	*	*	*	*	*	*																			
H2A.Z+H3K27ac	*	*	*	*	*	*	*	*	*	+															
H2A.Z+H3K9ac	*	*	*	*	*	*	*	*	*	+	+														
H3K4me1	*	*	*	*	*	*	*	*	*	*	*	*	+												
H3K4me3	*	*	*	*	*	*	*	*	*	*	*	*	*	*											
H3K9ac+H3K27ac	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
H2A.Z	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
H3K9ac	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	+					
H3K27ac	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	+	+					

Table 10. Friedman-Nemenyi hypothesis test results for the AUC metric and the models based on different histone modifications. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1. In this analysis, it was used data from the cell type **K562**.

	H3K4me1+H3K4me3+H3K9ac	H3K4me1+H3K4me3+H3K27ac	H3K4me3+H3K9ac+H3K27ac	H2A.Z+H3K4me1+H3K4me3	H2A.Z+H3K4me3+H3K9ac	H3K4me3+H3K9ac	H2A.Z+H3K4me3+H3K27ac	H3K4me1+H3K4me3	H3K4me1+H3K9ac+H3K27ac	H3K4me3+H3K27ac	H2A.Z+H3K4me1+H3K9ac	H2A.Z+H3K4me3	H3K4me1+H3K9ac	H2A.Z+H3K4me1+H3K27ac	H3K4me3	H3K4me1+H3K27ac	H2A.Z+H3K9ac+H3K27ac	H2A.Z+H3K4me1	H3K9ac+H3K27ac	H2A.Z+H3K9ac	H3K4me1	H3K9ac	H2A.Z+H3K27ac	H3K27ac	H2A.Z
H3K4me1+H3K4me3+H3K9ac																									
H3K4me1+H3K4me3+H3K27ac																									
H3K4me3+H3K9ac+H3K27ac																									
H2A.Z+H3K4me1+H3K4me3																									
H2A.Z+H3K4me3+H3K9ac																									
H3K4me3+H3K9ac																									
H2A.Z+H3K4me3+H3K27ac																									
H3K4me1+H3K4me3																									
H3K4me1+H3K9ac+H3K27ac																									
H3K4me3+H3K27ac																									
H2A.Z+H3K4me1+H3K9ac	*	*	*	+	+																				
H2A.Z+H3K4me3	*	*	*	*	*																				
H3K4me1+H3K9ac	*	*	*	*	*	+																			
H2A.Z+H3K4me1+H3K27ac	*	*	*	*	*	*	+																		
H3K4me3	*	*	*	*	*	*	*	*	*	+															
H3K4me1+H3K27ac	*	*	*	*	*	*	*	*	*	*															
H2A.Z+H3K9ac+H3K27ac	*	*	*	*	*	*	*	*	*	*	*	+	+												
H2A.Z+H3K4me1	*	*	*	*	*	*	*	*	*	*	*	*	*	*											
H3K9ac+H3K27ac	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*										
H2A.Z+H3K9ac	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	+								
H3K4me1	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*								
H3K9ac	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	+							
H2A.Z+H3K27ac	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
H3K27ac	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
H2A.Z	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*

Table 11. Intersection between footprints and ChIP-seq peaks (100 bps centered on peak's summit). It is shown the number of footprints which intersect at least one ChIP-seq peak (FP+ChIP), the number of ChIP-seq peaks which intersect at least one footprint and the total number of footprints.

Method	FP+ChIP	ChIP+FP	Total FPs
DH-HMM (3)	335890 (66.12%)	113058 (45.12%)	508000
Boyle	82576 (78.93%)	37396 (14.93%)	104624
Neph	372669 (74.73%)	71132 (28.39%)	498683

Table 12. Average footprints per DHS region regarding all DHS regions (All), promoter-proximal regions (Promoter), promoter-distal regions (Distal), regions with high co-binding occurrence (HOT) and regions with low co-binding occurrence (LOT).

Method	All	Promoter	Distal	HOT	LOT
DH-HMM (3)	4.6	7.31	4.11	4.54	1.32
Boyle	3.08	3.58	2.81	2.06	0.63
Neph	9.84	11.71	9.29	8.65	0.11

Table 13. Summary of DHS regions with and without any overlapping footprint predictions from segmentation-based methods.

Method	With	Without
DH-HMM (3)	110537 (98.67%)	1488 (1.33%)
Boyle	33990 (30.34%)	78035 (69.66%)
Neph	50660 (45.22%)	61365 (54.78%)

Table 14. Motifs enriched according to MEME-ChIP in DHS regions without footprints from DH-HMM with three histone modifications. It is shown the motif logo, MEME *e*-value, percentage of DHS regions that contained such motif (% Sites) and the putative TFs (determined with TOMTOM tool using Jaspar dataset). Only motifs with *e*-value > 0 are shown. Results are sorted increasingly by *e*-value.


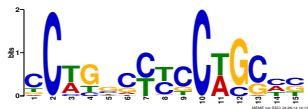
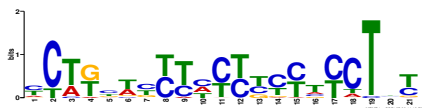

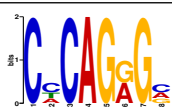
Motif Logo (MEME)	<i>e</i> -value (MEME)	% Sites (MEME)	Putative TFs (TOMTOM)
	8.5×10^{-193}	33.9%	ZNF263, Pax4, SP2, EGR1, KLF5, SP1, RREB1, EGR2
	3.9×10^{-23}	20.7%	No TF associated with this motif
	8.3×10^{-7}	6.3%	ZNF263, Erg
	3.8×10^{-5}	3.6%	EGR2
	5.4×10^{-5}	24.2%	TFAP2C, EBF1

Table 15. Friedman ranking for the AUC metric in different FIMO p -value thresholds when using Cuellar priors. For each metric, the methods are displayed in decreasing order with their respective Friedman ranking.

AUC Friedman Ranking	
$p\text{-value} = 10^{-5}$	1.2
$p\text{-value} = 10^{-6}$	2.3
$p\text{-value} = 10^{-4}$	2.8
$p\text{-value} = 10^{-7}$	4.15
$p\text{-value} = 10^{-3}$	4.55

Table 16. Friedman-Nemenyi hypothesis test results for the AUC metric in different FIMO p -value thresholds when using Cuellar priors. The asterisk and the cross, respectively, mean that the method in the column outperformed the method in the row with significance levels of 0.05 and 0.1

	$p\text{-value}=10^{-5}$	$p\text{-value}=10^{-6}$	$p\text{-value}=10^{-4}$	$p\text{-value}=10^{-7}$	$p\text{-value}=10^{-3}$
$p\text{-value}=10^{-5}$					
$p\text{-value}=10^{-6}$					
$p\text{-value}=10^{-4}$	*				
$p\text{-value}=10^{-7}$	*	*	+		
$p\text{-value}=10^{-3}$	*	*	*		

Table 17. Statistics for H1-hESC gold standard dataset.

Factor	ChIP-seq Peaks	Bit-score	MPBSs	ChIP+ MPBS(%)	Factor	ChIP-seq Peaks	Bit-score	MPBSs	ChIP+ MPBS(%)
ATF3	4804	13.2877 10.3789	86997 691899	24.81 37.03	BACH1	11457	13.2877 9.0247	73890 614421	7.87 25.7
CEBPB	15557	13.2877 10.3727	258034 1342548	28.95 62.49	CTCF	54070	13.2877 8.3074	65307 565933	47.41 77.68
EGR1	8743	13.2877 10.041	254088 1060314	42.06 59.93	GABPA	5652	13.2877 10.3874	23582 181503	24.15 38.32
C-jun	2148	13.2877 8.8895	149728 832374	15.64 30.17	JunD	9550	13.2877 8.5358	145422 717223	26.1 39.65
MAFK	11425	13.2877 10.0822	275246 1221488	46.56 68.71	MAX	11124	13.2877 11.1038	215060 855374	8.7 28.14
Myc	4551	13.2877 11.2326	204957 614797	11.01 25.55	NRF1	4513	13.2877 9.8346	29820 137117	66.76 80.57
POU5F1	3994	13.2877 9.4834	275838 2201678	47.97 69.15	RAD21	55674	13.2877 8.3074	65307 565933	46.04 76.64
REST	13269	13.2877 4.5733	26122 629168	31.44 48.56	RFX5	1695	13.2877 10.2372	131377 629248	30.03 41.12
RXRA	1306	13.2877 10.7089	350121 1110004	7.12 21.13	SIX5	3422	13.2877 9.8891	137900 1032447	32.09 49.09
SP1	15103	13.2877 10.6832	437639 1797400	21.55 35.35	SP2	2469	13.2877 10.2241	349740 1587339	30.82 50.67
SP4	5752	13.2877 9.4983	72668 503235	14.13 31.54	SRF	5102	13.2877 9.0972	152139 1024023	46.88 58.19
TCF12	7829	13.2877 11.0653	246510 893836	9.13 24.43	USF1	26028	13.2877 10.3789	86997 691899	38.6 70.27
USF2	6952	13.2877 10.5847	135124 759040	42.38 64.59	YY1	18310	13.2877 9.7387	186685 1325447	22.89 35.61
ZNF143	30687	13.2877 9.8891	137900 1032447	5.73 12.46					

Table 18. Statistics for K562 gold standard dataset (part 1).

Factor	ChIP-seq Peaks	Bit-score	MPBSs	ChIP+ MPBS(%)	Factor	ChIP-seq Peaks	Bit-score	MPBSs	ChIP+ MPBS(%)
ATF1	14864	13.2877 9.3778	20325 246442	6.53 17.6	ATF3	1233	13.2877 10.3578	– 496476	– 13.38
BACH1	3806	13.2877 9.0247	73890 614421	15.74 52.05	BHLHE40	22497	13.2877 10.9358	131233 572185	13.0 26.51
CCNT2	20057	13.2877 9.7947	121757 708983	6.58 11.46	CEBPB	38715	13.2877 10.3727	258034 1342548	35.61 64.05
CTCF	54387	13.2877 8.3074	65307 565933	46.13 75.62	CTCFL	11533	13.2877 8.3074	65307 565933	49.88 77.01
E2F4	8181	13.2877 10.4967	77280 173646	18.62 34.48	E2F6	16312	13.2877 10.7236	302788 1051116	13.55 26.16
eGFP-FOS	10256	13.2877 10.3019	202911 762222	65.78 85.93	eGFP-GATA2	11478	13.2877 10.5372	270225 1028569	12.83 33.66
EGR1	36997	13.2877 10.041	254088 1060314	41.29 68.08	eGFP-JunB	12287	13.2877 10.0929	210090 717235	46.44 63.47
eGFP-JunD	26674	13.2877 8.5358	145422 717223	25.62 41.4	ELF1	27780	13.2877 10.3812	229645 1026618	26.5 51.62
ELK1	2961	13.2877 10.4554	– 100691	– 44.48	ETS1	10726	13.2877 10.332	260115 1319961	5.99 16.24
C-fos	7646	13.2877 10.3019	202911 762222	34.17 44.79	FOSL1	11174	13.2877 9.4446	197170 699220	63.83 79.36
GABPA	14393	13.2877 10.3874	23582 181503	17.54 37.61	GATA1	4074	13.2877 10.4359	255049 1040470	21.13 47.32
GATA2	10648	13.2877 10.5372	270225 1028569	14.68 40.2	IRF1	8352	13.2877 7.4658	1199156 2330047	35.6 39.22
C-jun	9848	13.2877 8.8895	149728 832374	9.56 21.86	JunD	40052	13.2877 8.5358	145422 717223	23.38 38.47
MAFF	25074	13.2877 9.9826	195374 1215808	45.78 69.51	MAFK	19317	13.2877 10.0822	275246 1221488	41.64 64.33
MAX	31436	13.2877 11.1038	215060 855374	3.8 15.18	MEF2A	5631	13.2877 9.4253	820713 3210613	30.35 47.38

Table 19. Statistics for K562 gold standard dataset (part 2).

Factor	ChIP-seq Peaks	Bit-score	MPBSs	ChIP+ MPBS(%)	Factor	ChIP-seq Peaks	Bit-score	MPBSs	ChIP+ MPBS(%)
Myc	5023	13.2877	204957	11.15	NF-E2	2637	13.2877	75390	69.09
		11.2326	614797	26.16			7.5338	796063	82.56
NF-YA	4286	13.2877	79828	33.04	NF-YB	10096	13.2877	117371	53.79
		8.2746	428913	64.72			9.6997	470725	77.19
NR2F2	16678	13.2877	140437	5.87	NRF1	4211	13.2877	29820	59.18
		10.0106	626663	17.85			9.8346	137117	73.95
PU.1	28677	13.2877	432751	63.31	RAD21	17627	13.2877	65307	69.65
		9.9922	2040890	85.99			8.3074	565933	92.01
REST	15849	13.2877	26122	17.33	RFX5	2201	13.2877	131377	13.86
		4.5733	629168	26.46			10.2372	629248	21.58
SIX5	4194	13.2877	137900	23.56	SMC3	23598	13.2877	65307	65.28
		9.8891	1032447	37.08			8.3074	565933	87.97
SP1	7206	13.2877	437639	29.07	SP2	3124	13.2877	349740	35.53
		10.6832	1797400	45.49			10.2241	1587339	55.95
SRF	4717	13.2877	152139	22.01	STAT1	1476	13.2877	256939	5.96
		9.0972	1024023	31.27			9.1862	1272026	13.89
STAT2	1923	13.2877	370218	46.65	STAT5A	9811	13.2877	339560	15.34
		8.3974	3077582	58.92			10.4537	1292097	20.75
TAL1	26260	13.2877	121757	23.22	TR4	587	13.2877	101864	22.15
		9.7947	708983	43.25			8.8	825980	28.96
USF1	18521	13.2877	86997	36.68	USF2	3083	13.2877	135124	52.55
		10.3789	691899	64.62			10.5847	759040	73.79
YY1	4948	13.2877	186685	46.28	ZBTB33	3285	13.2877	12582	29.01
		9.7387	1325447	61.34			9.7537	82928	44.29
ZNF143	29069	13.2877	137900	5.9	ZNF263	3081	13.2877	2324743	35.18
		9.8891	1032447	12.55			9.1216	2577084	36.19

Table 20. Statistics for HeLa-S3 gold standard dataset.

Factor	ChIP-seq Peaks	Bit-score	MPBSs	ChIP+ MPBS(%)	Factor	ChIP-seq Peaks	Bit-score	MPBSs	ChIP+ MPBS(%)
CEBPB	61004	13.2877	258034	19.66	CTCF	52783	13.2877	65307	46.15
		10.3727	1342548	43.93			8.3074	565933	72.79
E2F4	2831	13.2877	77280	31.05	E2F6	4775	13.2877	302788	16.88
		10.4967	173646	49.45			10.7236	1051116	30.68
ELK1	4809	13.2877	–	–	C-fos	9325	13.2877	202911	52.75
		10.4554	100691	39.51			10.3019	762222	74.06
GABPA	6761	13.2877	23582	27.33	C-jun	21903	13.2877	149728	5.98
		10.3874	181503	52.83			8.8895	832374	15.15
JunD	31633	13.2877	145422	42.73	MAFK	14185	13.2877	275246	38.87
		8.5358	717223	67.01			10.0822	1221488	61.09
MAX	29647	13.2877	215060	2.84	Myc	10226	13.2877	204957	6.5
		11.1038	855374	10.83			11.2326	614797	16.15
NF-YA	5978	13.2877	79828	18.03	NF-YB	7156	13.2877	117371	33.02
		8.2746	428913	42.49			9.6997	470725	57.95
NRF1	2915	13.2877	29820	70.77	RAD21	43420	13.2877	65307	46.07
		9.8346	137117	81.27			8.3074	565933	70.0
REST	10247	13.2877	26122	30.96	STAT1	16158	13.2877	256939	23.13
		4.5733	629168	44.15			9.1862	1272026	35.09
USF2	12306	13.2877	135124	29.09	ZNF143	7048	13.2877	137900	14.78
		10.5847	759040	49.59			9.8891	1032447	26.53

Table 21. Statistics for HepG2 gold standard dataset.

Factor	ChIP-seq Peaks	Bit-score	MPBSs	ChIP+ MPBS(%)	Factor	ChIP-seq Peaks	Bit-score	MPBSs	ChIP+ MPBS(%)
BHLHE40	2859	13.2877	131233	22.46	CEBPB	18114	13.2877	258034	23.47
		10.9358	572185	41.55			10.3727	1342548	56.03
CTCF	55733	13.2877	65307	48.62	ELF1	17998	13.2877	229645	22.62
		8.3074	565933	79.54			10.3812	1026618	48.56
GABPA	10105	13.2877	23582	22.18	C-jun	12669	13.2877	149728	41.57
		10.3874	181503	46.79			8.8895	832374	56.37
JunD	21606	13.2877	145422	24.06	MAFF	37587	13.2877	195374	48.84
		8.5358	717223	39.35			9.9826	1215808	77.92
MAFK	61847	13.2877	275246	42.01	MAX	11852	13.2877	215060	3.81
		10.0822	1221488	71.65			11.1038	855374	17.77
Myc	4411	13.2877	204957	10.93	NRF1	1902	13.2877	29820	80.44
		11.2326	614797	26.34			9.8346	137117	89.33
RAD21	54261	13.2877	65307	46.27	REST	6021	13.2877	26122	35.61
		8.3074	565933	75.26			4.5733	629168	47.35
RXRA	17059	13.2877	350121	9.7	SP1	25465	13.2877	437639	11.16
		10.7089	1110004	27.26			10.6832	1797400	20.86
SP2	2626	13.2877	349740	11.42	SRF	5311	13.2877	152139	33.14
		10.2241	1587339	21.86			9.0972	1024023	50.71
USF1	21885	13.2877	86997	36.07	USF2	6290	13.2877	135124	51.65
		10.3789	691899	64.93			10.5847	759040	73.48
YY1	17871	13.2877	186685	13.69					
		9.7387	1325447	22.62					

Table 22. Summary of the DNase-seq and histone modification ChIP-seq data.

Cell Type	Data Type	Lab	UCSC Access.	GEO Access.	# Mapped Reads
H1-hESC	DNase-seq	Crawford	wgEncodeEH000556	GSM816632	110303078
H1-hESC	H3K4me1	Bernstein	wgEncodeEH000106	GSM733782	27286943
H1-hESC	H3K4me3	Bernstein	wgEncodeEH000086	GSM733657	19203931
H1-hESC	H3K9ac	Bernstein	wgEncodeEH000109	GSM733773	30288927
H1-hESC	H3K27ac	Bernstein	wgEncodeEH000997	GSM733718	31993560
H1-hESC	H2A.Z	Bernstein	wgEncodeEH002082	–	76761942
HeLa-S3	DNase-seq	Crawford	wgEncodeEH000540	GSM816643	54267867
HeLa-S3	H3K4me1	Bernstein	wgEncodeEH001750	GSM798322	38435440
HeLa-S3	H3K4me3	Bernstein	wgEncodeEH001017	GSM733682	35897578
HepG2	DNase-seq	Crawford	wgEncodeEH000537	GSM816662	50838536
HepG2	H3K4me1	Bernstein	wgEncodeEH001749	GSM798321	52320612
HepG2	H3K4me3	Bernstein	wgEncodeEH000095	GSM733737	18620773
K562	DNase-seq	Crawford	wgEncodeEH000530	–	365820647
K562	H3K4me1	Bernstein	wgEncodeEH000046	GSM733692	29197613
K562	H3K4me3	Bernstein	wgEncodeEH000048	GSM733680	25153055
K562	H3K9ac	Bernstein	wgEncodeEH000049	GSM733778	32634427
K562	H3K27ac	Bernstein	wgEncodeEH000043	GSM733656	24470196
K562	H2A.Z	Bernstein	wgEncodeEH001038	GSM733786	38763180

Table 23. Summary of the PWMs and TF ChIP-seq data for H1-hESC cell type.

Factor Name	ChIP-seq		PWM	
	Lab	UCSC Access.	Repository	PWM ID
ATF3	Myers	wgEncodeEH001566	Jaspar	MA0093.2
BACH1	Snyder	wgEncodeEH002842	Transfac	M00495
CEBPB	Snyder	wgEncodeEH002825	Jaspar	MA0466.1
CTCF	Myers	wgEncodeEH001649	Jaspar	MA0139.1
EGR1	Myers	wgEncodeEH001538	Jaspar	MA0162.2
GABPA	Myers	wgEncodeEH001534	Jaspar	MA0062.2
C-jun	Snyder	wgEncodeEH001854	Jaspar	MA0488.1
JunD	Snyder	wgEncodeEH002023	Jaspar	MA0491.1
MAFK	Snyder	wgEncodeEH002828	Jaspar	MA0496.1
MAX	Farnham	wgEncodeEH001757	Jaspar	MA0058.2
Myc	Snyder	wgEncodeEH002795	Jaspar	MA0147.2
NRF1	Snyder	wgEncodeEH001847	Jaspar	MA0506.1
POU5F1	Myers	wgEncodeEH001636	Jaspar	MA0142.1
RAD21	Snyder	wgEncodeEH001836	Jaspar	MA0139.1
REST	Myers	wgEncodeEH001498	Jaspar	MA0138.2
RFX5	Snyder	wgEncodeEH001835	Jaspar	MA0510.1
RXRA	Myers	wgEncodeEH001560	Jaspar	MA0512.1
SIX5	Myers	wgEncodeEH001528	Jaspar	MA0088.1
SP1	Myers	wgEncodeEH001529	Jaspar	MA0079.3
SP2	Myers	wgEncodeEH002302	Jaspar	MA0516.1
SP4	Myers	wgEncodeEH002317	Uniprobe	UP00002
SRF	Myers	wgEncodeEH001533	Jaspar	MA0083.2
TCF12	Myers	wgEncodeEH001531	Jaspar	MA0521.1
USF1	Myers	wgEncodeEH001532	Jaspar	MA0093.2
USF2	Snyder	wgEncodeEH001837	Jaspar	MA0526.1
YY1	Myers	wgEncodeEH001567	Jaspar	MA0095.2
ZNF143	Snyder	wgEncodeEH002802	Jaspar	MA0088.1

Table 24. Summary of the PWMs and TF ChIP-seq data for K562 cell type (part 1).

Factor Name	ChIP-seq		PWM	
	Lab	UCSC Access.	Repository	PWM ID
ATF1	Struhl	wgEncodeEH002865	Uniprobe	UP00020
ATF3	Struhl	wgEncodeEH000700	Jaspar	MA0018.2
BACH1	Snyder	wgEncodeEH002846	Transfac	M00495
BHLHE40	Snyder	wgEncodeEH001857	Jaspar	MA0464.1
CCNT2	Struhl	wgEncodeEH001864	Jaspar	MA0140.2
CEBPB	Snyder	wgEncodeEH001821	Jaspar	MA0466.1
CTCF	Snyder	wgEncodeEH002797	Jaspar	MA0139.1
CTCFL	Myers	wgEncodeEH001652	Jaspar	MA0139.1
E2F4	Farnham	wgEncodeEH000671	Jaspar	MA0470.1
E2F6	Farnham	wgEncodeEH000676	Jaspar	MA0471.1
eGFP-FOS	White	wgEncodeEH001207	Jaspar	MA0476.1
eGFP-GATA2	White	wgEncodeEH001208	Jaspar	MA0036.2
EGR1	Myers	wgEncodeEH001646	Jaspar	MA0162.2
eGFP-JunB	White	wgEncodeEH001210	Jaspar	MA0490.1
eGFP-JunD	White	wgEncodeEH001211	Jaspar	MA0491.1
ELF1	Myers	wgEncodeEH001619	Jaspar	MA0473.1
ELK1	Snyder	wgEncodeEH003356	Jaspar	MA0028.1
ETS1	Myers	wgEncodeEH001580	Jaspar	MA0098.2
C-fos	Snyder	wgEncodeEH000619	Jaspar	MA0476.1
FOSL1	Myers	wgEncodeEH001637	Jaspar	MA0477.1
GABPA	Myers	wgEncodeEH001604	Jaspar	MA0062.2
GATA1	Farnham	wgEncodeEH000638	Jaspar	MA0035.3
GATA2	Farnham	wgEncodeEH000683	Jaspar	MA0036.2
IRF1	Snyder	wgEncodeEH002798	Jaspar	MA0050.2
C-jun	Snyder	wgEncodeEH000620	Jaspar	MA0488.1
JunD	Snyder	wgEncodeEH002164	Jaspar	MA0491.1
MAFF	Snyder	wgEncodeEH002804	Jaspar	MA0495.1
MAFK	Snyder	wgEncodeEH001844	Jaspar	MA0496.1
MAX	Snyder	wgEncodeEH002869	Jaspar	MA0058.2

Table 25. Summary of the PWMs and TF ChIP-seq data for K562 cell type (part 2).

Factor Name	ChIP-seq		PWM	
	Lab	UCSC Access.	Repository	PWM ID
MEF2A	Myers	wgEncodeEH001663	Jaspar	MA0052.2
Myc	Snyder	wgEncodeEH000621	Jaspar	MA0147.2
NF-E2	Snyder	wgEncodeEH000624	Jaspar	MA0501.1
NF-YA	Snyder	wgEncodeEH002021	Jaspar	MA0060.2
NF-YB	Snyder	wgEncodeEH002024	Jaspar	MA0502.1
NR2F2	Myers	wgEncodeEH002382	Uniprobe	UP00009
NRF1	Snyder	wgEncodeEH001796	Jaspar	MA0506.1
PU.1	Myers	wgEncodeEH001482	Jaspar	MA0080.3
RAD21	Snyder	wgEncodeEH000649	Jaspar	MA0139.1
REST	Myers	wgEncodeEH001638	Jaspar	MA0138.2
RFX5	Snyder	wgEncodeEH002033	Jaspar	MA0510.1
SIX5	Myers	wgEncodeEH001483	Jaspar	MA0088.1
SMC3	Snyder	wgEncodeEH001845	Jaspar	MA0139.1
SP1	Myers	wgEncodeEH001578	Jaspar	MA0079.3
SP2	Myers	wgEncodeEH001653	Jaspar	MA0516.1
SRF	Myers	wgEncodeEH001600	Jaspar	MA0083.2
STAT1	Snyder	wgEncodeEH000664	Jaspar	MA0137.3
STAT2	Snyder	wgEncodeEH000666	Jaspar	MA0517.1
STAT5A	Myers	wgEncodeEH002347	Jaspar	MA0519.1
TAL1	Snyder	wgEncodeEH001824	Jaspar	MA0140.2
TR4	Farnham	wgEncodeEH000682	Jaspar	MA0504.1
USF1	Myers	wgEncodeEH001583	Jaspar	MA0093.2
USF2	Snyder	wgEncodeEH001797	Jaspar	MA0526.1
YY1	Farnham	wgEncodeEH000684	Jaspar	MA0095.2
ZBTB33	Myers	wgEncodeEH001569	Jaspar	MA0527.1
ZNF143	Snyder	wgEncodeEH002030	Jaspar	MA0088.1
ZNF263	Farnham	wgEncodeEH000630	Jaspar	MA0528.1

Table 26. Summary of the PWMs and TF ChIP-seq data for HeLa-S3 cell type.

Factor Name	ChIP-seq		PWM	
	Lab	UCSC Access.	Repository	PWM ID
CEBPB	Snyder	wgEncodeEH001815	Jaspar	MA0466.1
CTCF	Bernstein	wgEncodeEH001012	Jaspar	MA0139.1
E2F4	Snyder	wgEncodeEH000689	Jaspar	MA0470.1
E2F6	Snyder	wgEncodeEH000692	Jaspar	MA0471.1
ELK1	Snyder	wgEncodeEH002864	Jaspar	MA0028.1
C-fos	Snyder	wgEncodeEH000647	Jaspar	MA0476.1
GABPA	Myers	wgEncodeEH001504	Jaspar	MA0062.2
C-jun	Snyder	wgEncodeEH000746	Jaspar	MA0488.1
JunD	Snyder	wgEncodeEH000745	Jaspar	MA0491.1
MAFK	Snyder	wgEncodeEH002856	Jaspar	MA0496.1
MAX	Snyder	wgEncodeEH002830	Jaspar	MA0058.2
Myc	Snyder	wgEncodeEH000648	Jaspar	MA0147.2
NF-YA	Snyder	wgEncodeEH002066	Jaspar	MA0060.2
NF-YB	Snyder	wgEncodeEH002067	Jaspar	MA0502.1
NRF1	Snyder	wgEncodeEH000723	Jaspar	MA0506.1
RAD21	Snyder	wgEncodeEH001789	Jaspar	MA0139.1
REST	Myers	wgEncodeEH001629	Jaspar	MA0138.2
STAT1	Snyder	wgEncodeEH000614	Jaspar	MA0137.3
USF2	Snyder	wgEncodeEH001819	Jaspar	MA0526.1
ZNF143	Snyder	wgEncodeEH002028	Jaspar	MA0088.1

Table 27. Summary of the PWMs and TF ChIP-seq data for HepG2 cell type.

Factor Name	ChIP-seq		PWM	
	Lab	UCSC Access.	Repository	PWM ID
BHLHE40	Myers	wgEncodeEH001515	Jaspar	MA0464.1
CEBPB	Myers	wgEncodeEH002304	Jaspar	MA0466.1
CTCF	Myers	wgEncodeEH001516	Jaspar	MA0139.1
ELF1	Myers	wgEncodeEH001641	Jaspar	MA0473.1
GABPA	Myers	wgEncodeEH001548	Jaspar	MA0062.2
C-jun	Snyder	wgEncodeEH001794	Jaspar	MA0488.1
JunD	Myers	wgEncodeEH001470	Jaspar	MA0491.1
MAFF	Snyder	wgEncodeEH001841	Jaspar	MA0495.1
MAFK	Snyder	wgEncodeEH001842	Jaspar	MA0496.1
MAX	Snyder	wgEncodeEH002796	Jaspar	MA0058.2
Myc	Iyer	wgEncodeEH000545	Jaspar	MA0147.2
NRF1	Snyder	wgEncodeEH001802	Jaspar	MA0506.1
RAD21	Myers	wgEncodeEH001608	Jaspar	MA0139.1
REST	Myers	wgEncodeEH001549	Jaspar	MA0138.2
RXRA	Myers	wgEncodeEH001506	Jaspar	MA0512.1
SP1	Myers	wgEncodeEH001561	Jaspar	MA0079.3
SP2	Myers	wgEncodeEH002264	Jaspar	MA0516.1
SRF	Myers	wgEncodeEH001611	Jaspar	MA0083.2
USF1	Myers	wgEncodeEH001472	Jaspar	MA0093.2
USF2	Snyder	wgEncodeEH001804	Jaspar	MA0526.1
YY1	Myers	wgEncodeEH001661	Jaspar	MA0095.2