BIOINFORMATICS

Vol. 00 no. 00 2013
Pages 1–8

Detection of Active Transcription Factor Binding Sites with the Combination of DNase Hypersensitivity and Histone Modifications

Eduardo G. Gusmão [1], Christoph Dieterich [2], Martin Zenke [3,4] and Ivan G. Costa [1,5,6,*]

[1]IZKF Computational Biology Research Group, Institute for Biomedical Engineering, RWTH Aachen University Medical School, Germany.
[2]Computational RNA Biology and Ageing, Max Planck Institute for Biology of Ageing, Germany.
[3]Department of Cell Biology, Institute for Biomedical Engineering, RWTH Aachen University Medical School, Germany.
[4]Helmholtz Institute for Biomedical Engineering, RWTH Aachen University, Germany.
[5]Aachen Institute for Advanced Study in Computational Engineering Science (AICES), RWTH Aachen University, Germany.
[6]Center of Informatics, Federal University of Pernambuco, Brazil.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: The identification of active transcriptional regulatory elements is crucial to understand regulatory networks driving cellular processes such as cell development and the onset of diseases. It has recently been shown that chromatin structure information, such as DNase I hypersensitivity or histone modifications, significantly improves cell-specific predictions of transcription factor binding sites. However, no method have so far successfully combined both DNase I hypersensitivity and histone modification data to perform active binding site prediction.

Results: We propose here a method based on hidden Markov models to integrate DNase I hypersensitivity and histone modifications occupancy for the detection of open chromatin regions and active binding sites. We have created a framework which includes treatment of genomic signals, model training and genome-wide application. In a comparative analysis, our method obtained a good trade-off between sensitivity vs. specificity and superior area under the curve statistics than competing methods. Moreover, our technique does not require further training or sequence information in order to generate binding location predictions. Therefore, the method can be easily applied on new cell types and allow flexible downstream analysis such as *de novo* motif finding.

Availability: Our framework is available as part of the Regulatory Genomics Toolbox. The software information and all benchmarking data is available at http://costalab.org/wp/dh-hmm.

Contact: ivan.costa@rwth-aachen.de, eduardo.gusmao@rwth-aachen.de

*to whom correspondence should be addressed

1 INTRODUCTION

Transcriptional regulation orchestrates the proper temporal and spatial expression of genes (Maston, G. A. *et al.*, 2006). The identification of transcriptional regulatory elements, such as transcription factor binding sites (TFBSs), is crucial to understand regulatory networks driving cellular processes such as cell development and the onset of diseases. The standard approach is the use of sequence-based bioinformatics methods, which search over the genome for motif-predicted binding sites (MPBSs) representing the DNA binding sequence of a transcription factor (TF) (Stormo, 2000). This binding affinity is usually represented by models such as position weight matrices (PWMs). However, motifs are generally short and have low information content. Moreover, the presence of a sequence motif does not imply actual TF binding in that particular cell type. As a result, sequence-based methods return myriads of putative TFBSs, of which only a small fraction is functional in a particular context (Maston, G. A. *et al.*, 2006).

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) allows the detection of TFBSs *in vivo* on a genome-wide level (Landt, S. G. *et al.*, 2012). However, success of ChIP-seq assays depends on the existence of a good antibody against the TFs of interest and on the availability of large numbers of cells. These two conditions are not always met in particular for primary cells. Therefore, ChIP-seq based studies are restricted to the analysis of a small selection of TFs and cell types (Kim, J. *et al.*, 2008; Ouyang, Z. *et al.*, 2009) or require the effort of large consortia (ENCODE Project Consortium, 2012).

Another solution is to explore the fact that an 'open' chromatin structure is crucial and a prerequisite to the cell type-specific binding of a TF to DNA (Arvey, A. *et al.*, 2012; Thurman, R. E. *et al.*, 2012). For example, regions with histone modification marks H3K4me3 and H3K4me1 highlight active promoter and

enhancer elements (Bell, O. *et al.*, 2011). More detailed studies have demonstrated that active TFBSs occur between two regions with high active histone marks (peak-dip-peak pattern) (Hon, G. *et al.*, 2009). A classical method to detect active regulatory elements, DNase I footprinting, has also been recently combined with sequencing to indicate open chromatin regions on a genome-wide scale (Crawford, G. E. *et al.*, 2006). DNase-seq allows the identification of putative active TFBSs at base pair level resolution, i.e. 5–20 bp regions within two DNase-seq peaks represent the so-called footprints, where TFs are likely to be bound. While DNase-seq returns the exact position where DNA was nicked by DNase I enzyme, ChIP-seq returns sequences centered on the protein interacting with DNA. Therefore, DNase-seq gives a more precise estimate of open chromatin than histone modification ChIP-seq, as reflected in the sharpness of the signal of DNase-seq in comparison to H3K4me3 ChIP-seq signal (see Fig. 1A). A recent evaluation of DNase-seq on over 50 cell types indicated that DNase I digestion patterns describe the topology of protein–DNA binding interfaces and allowed the finding of 394 novel motifs (Neph, S. *et al.*, 2012).

Several computational approaches explore open chromatin evidence from ChIP-seq data or DNase I hypersensitivity (DHS) sites to restrict the search space of sequence-based analysis (Boyle, A. P. *et al.*, 2011; Cuellar-Partida, G. *et al.*, 2012; Gusmão, E. G. *et al.*, 2012; Natarajan, A. *et al.*, 2012; Neph, S. *et al.*, 2012; Pique-Regi, R. *et al.*, 2011; Whitington, T. *et al.*, 2009; Won, K. J. *et al.*, 2010). They show a clear advantage in detecting active TFBSs, as indicated by ChIP-seq evidence, in comparison to pure sequence-based methods.

1.1 Previous Approaches

Previous approaches can be categorized in two distinct classes: segmentation-based methods and site-centric methods. Segmentation-based methods such as Boyle, A. P. *et al.* (2011); Neph, S. *et al.* (2012); Won, K. J. *et al.* (2010), were based on the application of hidden Markov models (HMMs) or sliding window methods to segment the genome into open (and closed) chromatin region. Won, K. J. *et al.* (2010), for example, was based on the detection of typical peak-dip-peak profiles of H3K4me3 and H3K4me1 histone modifications indicative of active promoter and enhancer regions. Afterwards, Boyle, A. P. *et al.* (2011) proposed an HMM model to detect open chromatin/footprints in DHS signals on a base pair resolution. Lastly, Neph, S. *et al.* (2012) applies a simple sliding window approach that finds small regions (6–40 bp) with low DHS counts in the middle of regions (3–10 bp) with high DHS counts. Site-centric methods, on the other hand, perform predictions around MPBSs. For example, Pique-Regi, R. *et al.* (2011) first detects MPBSs then it applies an unsupervised learning method that uses chromatin information to group sites as active or inactive. This Bayesian method combines DHS profiles and histone modification summary statistics around the TFBSs with priors derived from motif matching bit-score, sequence conservation and distance to the nearest transcription start site. Later, Cuellar-Partida, G. *et al.* (2012) improved the methodology proposed in Whitington, T. *et al.* (2009) to include the combination of DHS and distinct histone modification as priors in the detection of active MPBSs. They showed that DHS improves TFBS detection considerably. Nevertheless, no significant improvement was observed when DHS was combined with histone modifications. Interestingly, Pique-Regi, R. *et al.* (2011) also could not observe advantages in using histone modifications. Note, however, that both Pique-Regi, R. *et al.* (2011) and Cuellar-Partida, G. *et al.* (2012) use simple window-based statistics for histone modification, which are unable to detect peak-dip-peak shapes as in Won, K. J. *et al.* (2010).

1.2 Our Approach

We propose here an HMM-based approach to integrate both DHS (DNase-seq) and histone modifications (ChIP-seq) for the detection of open chromatin regions and active TFBSs. We and others have previously observed that the peak-dip-peak patterns of the DHS profile happen inside the dip of the histone modification profiles (Gusmão, E. G. *et al.*, 2012; Kundaje, A. *et al.*, 2012) (see Fig. 1A). This indicates an underlying epigenetic grammar behind active TFBSs: open chromatin regions (high DHS) are flanked by high or moderate signals of active histone marks. These open chromatin regions consist of a sequence of one or more footprints indicative of active TFBSs (low histone modification and DHS signals). We have therefore devised an HMM (Fig. 1B) to model this epigenetic grammar by simultaneous analysis of DNase-seq and the ChIP-seq profiles of the histone marks H3K4me1, H3K4me3, H3K9ac, H3K27ac and H2A.Z, which are indicative of active regulatory regions, on a genome-wide level. The HMM has as input a normalized and a slope signal of DHS and one histone mark. It can therefore detect the increase, top and decrease regions of both histone mark and DHS signals. The HMM has multivariate Gaussian density functions with full covariance matrices as emission functions in order to capture correlations between the DHS and histone marks signals. As in Boyle, A. P. *et al.* (2011), we trained the HMM on the annotation of a single genomic region. Moreover, we devised a normalization procedure of the DHS/histone mark profiles to allow the application of an HMM trained in a particular cell type to be applied on any cell type of interest. This is the first approach combining local genomic profiles of histone modification and DHS for the detection of open chromatin and active TFBSs. We evaluate our and competing methods with public data from H1-hESC and K562 cell types. To validate our predictions, we created datasets by combining MPBSs with ChIP-seq from 83 TFs.

2 METHOD

2.1 Epigenetic Signal Processing

Let the matrix $\mathbf{X}$ representing genomic signals be defined as

$$\mathbf{X} = \{x_{ij}\}^{D \times N},$$

where D is the number of genomic signals and N refers to the number of bases in the genome. The ith genomic signal is represented by the vector

$$\mathbf{x_{i\cdot}} = \{x_{i1}, ..., x_{iN}\}.$$

and the genomic signals at the jth position are represented as

$$\mathbf{x_{\cdot j}} = \{x_{1j}, ..., x_{Dj}\}.$$

The first processing step consists on creating a base-pair resolution genomic coverage signal by counting the reads mapped to the genome. For DHS data, we only consider the first 5′ position of the aligned reads (corresponding to the exact position at which DNase I enzyme has nicked the DNA). For histone modification ChIP-seq data, we extend all the aligned

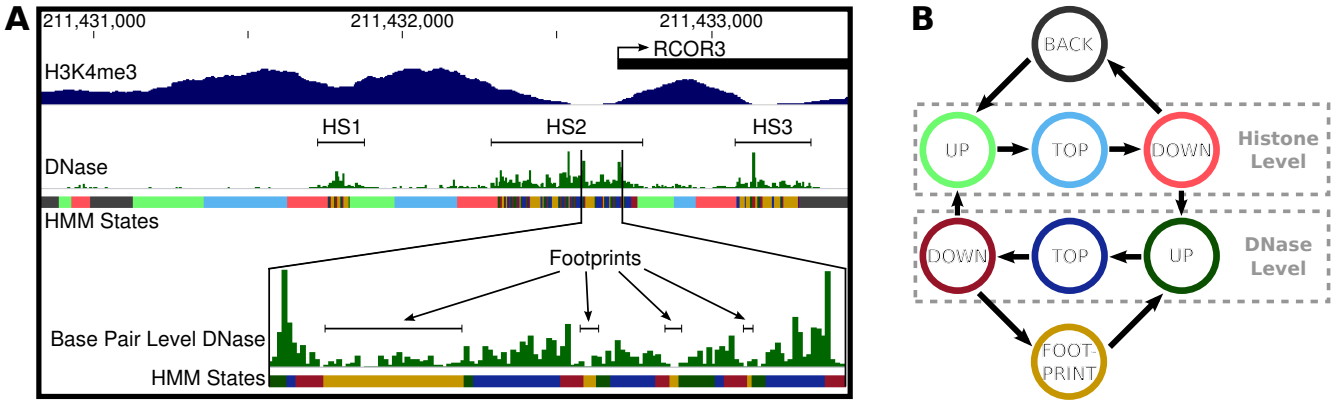

Fig. 1. Epigenetic grammar. (**A**) DHS (DNase-seq) and H3K4me3 (ChIP-seq) profiles around the promoter region of RCOR3 – REST corepressor 3 on K562 cell type. The DNase-seq signal indicates three clear regions of DHS (HS1, HS2 and HS3), each of which fits dip regions within the H3K4me3 signal. Moreover, these regions consist of several putative footprints of varied sizes. (**B**) Eight state HMM proposed. The first state models background signal (`BACK`). From the background state, the only possible transition is to the histone level states, which will model the increase (`UP`), high levels (`TOP`) and decrease (`DOWN`) of histone modifications. After visiting histone level states, the HMM allows transitions to the DNase level states, which again model the increase, high levels and decrease in the DHS signal. Only then, the `FOOTPRINT` state can be visited. After a footprint visit, the HMM has to go again to the DNase and histone level states, emphasizing the peak-dip-peak pattern. We omit the self-transitions, which are present in all states, for simplicity.

reads to be 200 bp long, as this is the expected length of immunoprecipitated fragments. Then, in both cases, the resulting genomic signal is created by counting how many reads overlapped at each genomic position.

The signals are then normalized using an approach that addresses both within- and between-dataset variability. For that, the genome is partitioned into a set $\{b_1, ..., b_M\}$ of non-overlapping bins and a set $\{r_1, ..., r_M\}$ of overlapping bins. Each bin b_m covers the genomic coordinate interval $[((m-1)\cdot L)+1, m\cdot L]$, and r_m represents b_m extended by $L/2$ for both sides. By using $L = 5000$, we partition the genome in regions with total length 10000. First, we apply a within-signal normalization by averaging non-zero read counts inside bins (Boyle, A. P. *et al.*, 2011). For a given position x_{ij}, such that $j \in b_m$, we apply

$$x_{ij}^{norm1} = \frac{x_{ij}}{\sum\limits_{w \in r_m} x_{iw}\mathbf{1}(x_{iw} > 0) \Big/ \sum\limits_{w \in r_m} \mathbf{1}(x_{iw} > 0)}, \quad (1)$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. Next, we perform a between-dataset scaling procedure following (Hon, G. *et al.*, 2009) to force values inside the interval $[0, 1]$. Assuming $j \in b_m$, this is done by applying a logistic function to the first normalized values as follows

$$x_{ij}^{norm2} = \frac{1}{1 + e^{-(x_{ij}^{norm1} - P_{r_m}^t)/\sigma_{r_m}}}, \quad (2)$$

where $\sigma_{r_m} = \sqrt{\sum_{w \in r_m}(x_{iw} - \mu_{r_m})^2/(2L)}$, $\mu_{r_m} = \sum_{w \in r_m} x_{iw}^{norm1}/(2L)$ and $P_{r_m}^t$ are, respectively, the standard deviation, mean and the tth percentile of values $x_{iw}^{norm1} \in r_m$.

In order to estimate the slope of the signals we apply a Savitzky-Golay smoothing filter. This method consists of fitting the data into a 2nd order polynomial, performing a convolution (based on a specific window length) with a vector containing Savitzky-Golay coefficients (Madden, 1978). The window length was set to 9 bp (including the central element) for DHS data (Boyle, A. P. *et al.*, 2011) and 201 bp for histone modification ChIP-seq data, matching the read extension length. Next, the first derivative is applied to the smoothed signal. The resulting signal represents the slope of the normalized curve and assumes positive values when there is an increase and negative values when there is a decrease. The slope signal will help the delineation of the start and end of DHS and histone modification peaks (see Supplementary Fig. 2 for examples of these signals).

2.2 Prediction of footprints with HMMs

The HMM structure was designed to recognize the epigenetic grammar described in Section 1.2. This segmentation task is performed based on four input signals: the normalized and slope versions of both DHS and a histone modification. The structure, depicted in Fig. 1B can be interpreted as follows. The first state (`BACK`) corresponds to the 'background' regions with low concentration of DHS and histone marks. The histone level states represent a peak in the histone modification signal, recognizing an increase in the histone modification signal based on high positive slope values (`UP`), summit regions with slope values close to zero with high normalized values (`TOP`) and a decrease based on negative values of the slope signal (`DOWN`). From the histone level `DOWN` state, the model can either return to `BACK` (isolated histone modification peaks without further DHS) or continue to the DNase level `UP` state. The DNase level states are equivalent to the histone level states, with the exception that the DHS normalized and slope signals are being recognized instead. From the DNase level `DOWN` state, the model decides between returning to a region of higher histone modification signals (histone level `UP` state) and visiting the `FOOTPRINT` state, which represents the dip between two peaks of intense DHS. The regions of the genome where the HMM has recognized as `FOOTPRINT` are the ones reported by our method as likely TFBSs. See Supplementary Section 3.3 for alternative HMM topologies.

More formally, for an observed multivariate sequence $\mathbf{X}$ and a hidden variable $\mathbf{Q} = \{q_1, ..., q_N\}$, we can describe an HMM by parameters $\Theta = A, B, \Pi\}$. A represents the state transition matrix

$$A = \{a_{uv}\}^{S \times S},$$

where S is the number of states and a_{uv} represents the probability of transition from state u to v. The initial state transition probabilities are represented as

$$\Pi = \{\pi_1, ..., \pi_S\}.$$

We use a multivariate normal density function with full covariance matrix as emission probability of the states. For a given genomic position j and state u we have

$$\begin{aligned} p(\mathbf{x}_{\cdot\mathbf{j}}|q_j = u) &= p(\mathbf{x}_{\cdot\mathbf{j}}|\mu^u, \Sigma^u) \\ &= \frac{1}{\sqrt{(2\pi)^D|\Sigma^u|}} e^{-\frac{1}{2}(\mathbf{x}_{\cdot\mathbf{j}} - \mu^u)^T(\Sigma^u)^{-1}(\mathbf{x}_{\cdot\mathbf{j}} - \mu^u)}, \end{aligned} \quad (3)$$

where μ^u and Σ^u are respectively the D-dimensional mean vector and full covariance matrix. Lastly, the emission parameters B are defined as $\{(\mu^1, \Sigma^1), ..., (\mu^S, \Sigma^S)\}$.

2.2.1 Model Training and Decoding The HMM is trained with maximum likelihood parameters on a supervised approach. For a given annotation sequence of the hidden data $\mathbf{Q}$ and sample data $\mathbf{X}$, the parameters are estimated as

$$a_{uv} = \frac{\alpha'_{uv}}{\sum_{w=1}^{S} \alpha'_{uw}}, \tag{4}$$

where $\alpha'_{uv} = \sum_{i}^{N-1} \mathbf{1}(q_i = u, q_{i+1} = v)$ represents the number of transitions from state u to state v observed in the training data.

As we expect the HMM to always start at the `BACK` state, the initial transition vector $\pi_1 = 1$ and $\pi_u = 0$ for $u > 1$.

Finally, the emission parameters are estimated as

$$\mu^u = \frac{\sum_{j=1}^{N} x_{\cdot j} \mathbf{1}(q_j = u)}{\sum_{j=1}^{N} \mathbf{1}(q_j = u)}, \tag{5}$$

and

$$\Sigma^u = \frac{\sum_{j=1}^{N} (x_{\cdot j} - \mu^u)^T (x_{\cdot j} - \mu^u) \mathbf{1}(q_j = u)}{\sum_{j=1}^{N} \mathbf{1}(q_j = u) - 1}. \tag{6}$$

The final goal of our model is to find the most probable states visited for a given genomic signal. More formally, we want to find the most probable sequence of hidden states $\mathbf{Q}$ having observed $\mathbf{X}$ given a model Θ, which can be written as

$$\mathbf{Q}^* = \arg\max_{\mathbf{Q}} p(\mathbf{X}, \mathbf{Q}|\Theta). \tag{7}$$

The solution to the above problem is given by the Viterbi algorithm (Rabiner, 1989). In practical terms, we consider positions annotated with the `FOOTPRINT` state to be potential active TFBSs.

3 EXPERIMENTAL DESIGN

3.1 Datasets

Both DNase-seq and ChIP-seq for TFs and histone modifications data were obtained in ENCODE repository (ENCODE Project Consortium, 2012). We used read alignments available for the embryonic stem cell (H1-hESC) and myelogenous leukemia (K562). DNase-seq and ChIP-seq for H3K4me1, H3K4me3, H3K9ac, H3K27ac and H2A.Z were downloaded for every cell type in order to generate the input for the HMMs (see Supplementary Table 22 for details on input data).

In order to create the evaluation dataset, we obtained all ChIP-seq enriched regions for TFs of these two cell types in ENCODE Analysis Working Group (AWG) data track. Also, we used PWMs from Jaspar (Mathelier, A. *et al.*, 2014), Transfac (Matys, V. *et al.*, 2006) and Uniproble (Robasky and Bulyk, 2011) repositories (see Supplementary Tables 23–25 for details on evaluation data).

All experimental files and alignments are based on the human genome build 37 (hg19). Chromosome Y has been removed from all analyses.

3.2 HMM Training and Evaluation

To reduce the dimensionality of the data, we first applied a peak calling tool to find regions with evidence of DHS and histone modification signals. The enriched regions of histone modifications ChIP-seq data were defined using the peak-calling tool MACS (Zhang, Y. *et al.*, 2008). We used a p-value of 10^{-5} and all default parameters from MACS 1.4. No further filtering, such as false discovery rate, was performed on the peaks, as we wanted a lenient selection of candidate regions. The enriched regions of the DHS data were defined as in Boyle, A. P. *et al.* (2011). Briefly, a signal corresponding to the estimated density of the DHS is generated by applying F-seq software (Boyle, A. P. *et al.*, 2008) to the DNase-seq mapped reads and background information based on alignability, copy number and karyotype correction. A threshold is then calculated by fitting the signal to a gamma distribution and considering the value that corresponds to a loose p-value of 0.01. We merge all enriched regions for a given cell type and extend them by 5000 bp in each direction. This step keeps only $3 - 6\%$ of the genome with DHS or histone modification evidence for a given cell type (see Supplementary Table 5 for complete statistics).

We selected a 10,000 bp region around the promoter of the gene RCOR3 and performed a cell type-specific manual annotation with one of the 8 HMM states according to the epigenetic grammar described in Fig. 1. As one of the histones marks – H3K4me1 – is known to be associated to distal enhancers, we have additionally annotated an enhancer region. The selection of these regions was made randomly, but we checked ENCODE tracks for evidence that the gene RCOR3 was expressed in all cell types analyzed and that the enhancer region was far (>100kb) from known genes and expressed regions.

In order to help the annotation of the footprints, MPBSs with all PWMs from Jaspar, Transfac and Uniprobe datasets were detected inside the training regions. We consider (active) footprints all the signal depleted regions between two DHS peaks that overlap a MPBS. We trained five HMMs per cell type, one for each histone modification (H3K4me3, H3K9ac, H3K29ac and H2A.Z with the promoter region and H3K4me1 with the enhancer region). The regions used for training were excluded from all further predictions. We used the Viterbi algorithm to find the footprint regions throughout the genome for each trained HMM. Note that the evaluation of the HMM on the genomic signals was performed regardless of any evidence that the regions are distal/proximal to a gene. See Supplementary Section 3.1 for details on training and Supplementary Tables 2–4 for the set of parameters for the HMM trained using DHS+H3K4me3 on H1-hESC data.

3.3 Evaluation

3.3.1 Binding Evidence ChIP-seq experiments for the TFs being tested were used as experimental evidence of binding. These are simply the enriched regions (or peaks) based on the uniform processing of ENCODE Analysis Working Group (AWG) data track. Next, we used the *de novo* motif analysis from Factorbook (Wang, J. *et al.*, 2013) to select the PWMs associated with the TF ChIP-seq data. We considered only the TFs in which the top enriched motif was present in at least 300 of the 500 top-scored ChIP-seq peaks. We used PWMs from the Jaspar database matching those in Factorbook. Four PWMs (ATF1, BACH1, NR2F2 and SP4) not present in Jaspar were obtained from Transfac and Uniprobre.

These PWMs were matched against the complete human genome using the motif matching tool available in Biopython (Cock, P. J. A. *et al.*, 2009). Initially, a regularizing value of 0.05 was added for all nucleotides at all positions of the PWMs. We used a false positive rate (FPR) (10^{-4}) approach based on dynamic programming (Wilczynski, B. *et al.*, 2009) to detect significant MPBSs. In this scenario, a different threshold is calculated for each PWM by defining the bit-score that corresponds to a specific FPR in the distribution of scores of that TF's PWM. Note that Pique-Regi, R. *et al.* (2011) used a fixed bit-score cutoff of $log_2(10000) = 13.288$ for all PWMs. However, we observed that this criterion is very strict, in the sense that only a few MPBSs coincide with regions enriched with experimental ChIP-seq data (see Supplementary Section 2.2 for discussion). Finally, we filtered out all TFs in which less than 10% of ChIP-seq peaks contained at least one MPBS associated. This resulted in a set of 56 TFs from K562 and 27 TFs for H1-hESC cell types (see Supplementary Tables 23–25 for the final selection of PWMs and TFs).

3.3.2 Gold Standard and Evaluation Metrics To evaluate all methods, a site-centric gold standard was proposed (Cuellar-Partida, G. *et al.*, 2012). In this evaluation scheme MPBSs with ChIP-seq evidence (i.e. lying within 100 bp from the peak summit) are considered 'true' TFBSs and all other MPBSs are considered 'false' TFBSs. Then, every footprint prediction that overlaps by at least 1 bp with a true TFBS is considered

a correct prediction (true positive – TP) and every footprint that overlaps with a false TFBS is considered an incorrect prediction (false positive – FP). Consequently, true negatives (TN) and false negatives (FN) are, respectively, false and true TFBSs without overlapping footprint predictions. With such contingency table we are able to calculate the sensitivity and specificity of each method. In Supplementary Tables 17–19 we show statistics on the number of MPBSs, ChIP-seq peaks and combinations of both.

To access the sensitivity $TP/(TP + FN)$ vs. specificity $TN/(TN + FP)$ trade-off we created receiver operating characteristic (ROC-like) curves (with true negative rate (specificity) on x-axis, instead of the traditional false positive rate) and estimated the area under the ROC-like curves (AUCs) as follows. For each method, the MPBSs from the gold standard were divided into two groups: the ones that contain at least 1 bp overlap with the predicted sites and the ones that do not overlap. Both groups were sorted based on the motif matching bit-score. A single list is then obtained by combining the ranked list of predicted sites before the ranked list of the non-predicted sites. A ROC-like curve was evaluated based on this list, for all cell types and TFs.

3.4 Competing Methods

We compared our method with Boyle method (Boyle, A. P. *et al.*, 2011), Centipede (Pique-Regi, R. *et al.*, 2011), Cuellar-Partida method (Cuellar-Partida, G. *et al.*, 2012) and Neph method (Neph, S. *et al.*, 2012) using the evaluation methodology defined previously. Boyle and Neph methods analyses were based on results/parameterization available in the original studies. As Cuellar-Partida used a distinct statistical framework for PWM detection, we performed experiments to select the cutoff criteria. Centipede presented very poor results with default parameters. We have therefore performed a grid search strategy to detect the best parameterization in one cell type and applied to the other cell type. This leads to results close to an optimistic parameterization, where parameter tuning was performed for each TF and cell type combination. Note that this optimistic parameterization is only possible with ChIP-seq for every TF tested. Moreover, the estimated parameters from both cell types were similar (level of shrinkage of negative binomial's parameters was estimated as 0.0 for H1-hESC and 0.25 for K562 and level of shrinkage of multinomial's parameters was estimated as 0.75 for both cell types) and represent a better choice of default parameters than the ones provided by the tool. See Supplementary Section 4 for further details on parameterization experiments for the competing methods. As we could not succeed in running the HMM-based histone segmentation approach from Won, K. J. *et al.* (2010), we adapted our approach to use only histone modification signals. For such, the DNase level states were replaced by a single `FOOTPRINT` state. All the steps described in Section 3.2 were performed again using new manually annotated regions based on the new HMM state configuration. This method will be referenced as H-HMM (Histone HMMs). All parameter tuning experiments were performed on chromosome 1 and these regions were excluded from the method comparison analysis. All annotations, predictions and evaluation data are available in our web supplement for future benchmarking purposes.

3.5 HMM Parameter Selection

We have performed a set of experiments to evaluate/justify methodological choices for our approach. For this, we used the genomic signals of chromosome 1, which was left out of any further analysis. First, we evaluated choices of scaling parameters: use of global or local statistics in Eq. 2 and value of the percentile (96%, 98% and 99%). Results indicate the advantage of local normalization and that 98% was a good trade-off between sensitivity and specificity. We have also compared the use of the Viterbi algorithm and posterior decoding to detect the footprints. Experimental results indicate a slight advantage of the Viterbi algorithm, while the posterior decoding had numerical problems in particular genomic regions. We have also evaluated two distinct HMM topologies. The first alternative merges `UP`–`TOP`–`DOWN` states in one to obtain a simple HMM. This HMM has very poor performance, as it does not take advantage of the slope of the DHS/histone modification signals. We have also extended the original HMM topology by including transitions from the background states directly to the DNase level states. This modification allows the detection of DHS peaks between asymmetric histone modification signals, which were evidenced in Kundaje, A. *et al.* (2012). The HMM had smaller AUC values than the HMM model proposed here and was therefore not explored. See Supplementary Section 3 for further discussion regarding all empirical analyses on HMM parameter selection. All HMMs were implemented using Scikit (Pedregosa, F. *et al.*, 2011). All experiments were executed in 4 Xeon E7-4870 CPUs with 10 2.4GHz cores each.

4 RESULTS

4.1 Selection of Histone Modifications

Given the predictions made by our method (referenced from this point as DH-HMM – DHS+Histone HMMs) in which different histone modifications are used as input, it is possible to create combined footprints. Briefly, we take the union of all predictions from any number of models and merge all overlapping footprints. Then, an initial question would be the selection of the optimal set of histone marks to be combined. We are particularly interested in using as few marks as possible to minimize the necessity of high-throughput experiments. For such, we have evaluated the AUC values of all combinations (up to three) of the five histone marks on chromosome 1. Results indicate that combinations with more histone marks are better than single-histone models (see Supplementary Fig. 7 and Table 7). Several combinations of three marks (H3K4me1+H3K4me3+H3K9ac, H3K4me1+H3K4me3+H3K27ac, H2A.Z+H3K4me1+H3K4me3, H2A.Z+H3K4me3+H3K9ac and H3K4me3+H3K9ac+H3K27ac) were similarly good, i.e. their AUC are not significantly lower than any other combination). Similarly, if we only consider individual and pairs of histone marks, H3K4me1+H3K4me3, H3K4me3+H3K9ac, H3K4me3+H3K27ac, H2A.Z+H3K4me3 and H3K4me1+H3K9ac have similar AUCs. This indicates that any combination of these histone marks, whenever available, would perform equally well. We have selected the combinations H3K4me1+H3K4me3+H3K9ac and H3K4me1+H3K4me3, which we call DH-HMM(3) and DH-HMM(2), respectively, for further analysis. We also used the same histone modification selection for all competing methods that use histone modification evidence (Cuellar(2), Cuellar(3), H-HMM(2) and H-HMM(3)). Moreover, we have performed this analysis for H1-hESC and K562 cell types in separate. Despite small differences in rankings, a similar set of combinations resulted as equally good and both H3K4me1+H3K4me3+H3K9ac and H3K4me1+H3K4me3 were among the top 2 combinations in either cell. See Supplementary Section 3.5 for further discussions.

4.2 Cell type specific vs. non-cell type specific training

Next, we have analyzed if the DH-HMM models are specific for the cell types they are trained on. In this particular test, we have extended our datasets to include two new cell types (HeLa-S3 and HepG2) with 20 and 21 TFs, respectively (more details on Supplementary Tables 20–21 and 26–27). We have compared the AUC values of the H3K4me1+H3K4me3 DH-HMM when it was trained in a particular cell type and executed in the same cell type vs. the other three cell types. A statistical test (paired Mann-Whitney-Wilcoxon with null hypothesis that the distributions are equal) was performed and showed that only in 1 out of 12 comparisons the cell

type-specific training had significantly superior AUC than a non-specific training (see Supplementary Section 3.6). This indicates that the DH-HMM models can be applied to any other cell type data with an insignificant loss of performance.

4.3 Method Comparison

We show in Fig. 2 the ROC-like curves for TFs GABPA, C-jun, SIX5 and YY1 and cell types H1-hESC and K562. All methods analyzed provided a list of active TFBSs given a particular parameterization (see Supplementary Section 4 for details). An exception is Centipede, in which we used the model's posterior probability to rank sites and the suggested probability (0.99) to select active TFBSs as performed in Pique-Regi, R. *et al.* (2011). The curves were obtained by ranking the active TFBSs predictions in regard to the PWM bit-score. To obtain complete curves, we also included in the end of the ranking the inactive TFBSs (sorted by PWM bit-scores). Squares (and circles) in the curve indicate the location of the rank with TFBSs predicted to be active only. These points were used to obtain the sensitivity and specificity of a method.

We tested the ranks of the methods for all 83 combinations of cell types and TFs regarding AUC, sensitivity and specificity. We used the Friedmann-Nemenyi test to evaluate the statistical significance of the method's ranks. As indicated in Table 1 and Supplementary Table 1, HMMs based on histone modification only (H-HMM) had significantly lower values for all indices. This can be explained by the fact that the histone modification signal, measured by ChIP-seq, contains a much lower resolution than, for instance, the signal obtained with DNase-seq, which is used by all other competing methods.

Boyle and Neph methods have significantly higher specificity values than competing methods (Table 1). We observed that, in general, Boyle makes very few predictions, resulting in very few false positives, but missing many of the true active TFBSs. For example, from the 3020 observed active GABPA binding sites in H1-hESC, only 2066 (68.41%) and 2207 (73.08%) were detected by Boyle and Neph, respectively; while Centipede and the DH-HMM(3) predicted 2765 (91.56%) and 2892 (95.76%), respectively. Indeed, Boyle and Neph methods' sensitivity is significantly lower than all other DHS-based methods.

The Cuellar method with either 2 or 3 histone modifications presented the highest sensitivity values, which were statistically higher than all methods but Centipede and DH-HMM(3) (Table 1). On the other hand, Cuellar had very poor specificity values being significantly lower values than all DHS-based methods. Since the number of false TFBSs is very high, Cuellar usually predicts a great number of false positives. For instance, Cuellar(3) predicts 5.63% (10,051 sites) more false positives than DH-HMM(3) on GABPA binding in H1-hESC. DH-HMM(2) and DH-HMM(3) significantly outperformed all other methods concerning AUC values (Table 1) and were significantly better ranked than all other methods in Friedman ranking (Supplementary Table 1).

4.4 Spatial Specificity and DHS Coverage of Segmentation Approaches

Next, we evaluated the spatial specificity (Wilbanks and Facciotti, 2010), i.e. the distance of all predicted regions to the center of their recognized active TFBSs. Note that this comparison is only made on segmentation-based methods (Boyle, Neph and HMM-based methods), as site-centric methods always have an 'ideal' spatial specificity since they use sequence information. The Fig. 3 shows the resulting distribution of distances for all evaluated TFs. Overall, DH-HMMs had better spatial specificity (lowest distance of footprints to the center of active TFBSs) than all other methods (Mann-Whitney-Wilcoxon of equal distance distributions was rejected with p-value $\leq 10^{-5}$). On the other hand, the H-HMMs presented large distances than all other methods (p-value $\leq 10^{-5}$). This is again explained by the lower resolution of the ChIP-seq to indicate chromatin structure in comparison to the DHS signal. These results indicate that DH-HMMs improve Boyle and Neph methods upon the detection of the exact location of TFBSs. Lastly, we evaluated footprints statistics inside DHS sites. Footprints from DH-HMM cover 98.67% of DHS sites, while footprints predicted by Boyle and Neph method cover only 30.34% and 45.22% of DHS sites. An inspection of the DNase-seq read coverage indicates that both Boyle and Neph methods fail to identify footprints in DHS sites with low to average read counts. See Supplementary Section 3.8 for further details.

Table 1. Friedman-Nemenyi hypothesis test results on sensitivity, specificity and AUC. 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.

		DH-HMM (3)	DH-HMM (2)	Centipede	Neph	Cuellar (2)	Cuellar (3)	H-HMM (3)	H-HMM (2)	Boyle
Sensitivity	Cuellar (2)									
	Cuellar (3)									
	DH-HMM (3)									
	Centipede									
	DH-HMM (2)					*	*			
	H-HMM (3)	+				*	*			
	H-HMM (2)	*	*	*		*	*	*		
	Neph	*	*	*		*	*	*		
	Boyle	*	*	*		*	*	*	*	
Specificity	Boyle									
	Neph									
	DH-HMM (2)				+					*
	Centipede				*					*
	DH-HMM (3)				*					*
	Cuellar (2)	*	*	*	*					*
	H-HMM (2)	*	*	*	*					*
	Cuellar (3)	*	*	*	*					*
	H-HMM (3)	*	*	*	*	*			+	*
AUC	DH-HMM (3)									
	DH-HMM (2)									
	Centipede	*	*							
	Neph	*	*							
	Cuellar (2)	*	*							
	Cuellar (3)	*	*							
	H-HMM (3)	*	*	*	*	+				
	H-HMM (2)	*	*	*	*	*	*			
	Boyle	*	*	*	*	*	*			

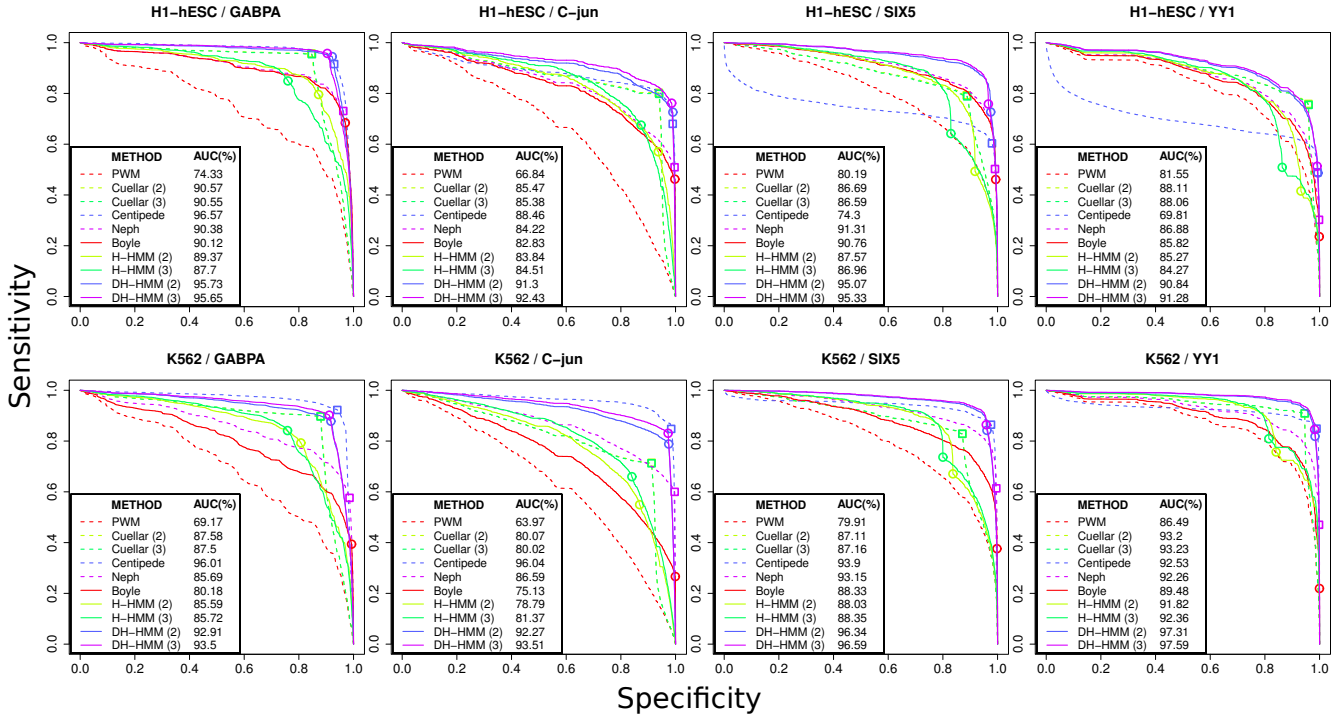

Fig. 2. ROC-like curves for a selection of TFs created when applying the methods to data from the cell types H1-hESC and K562.

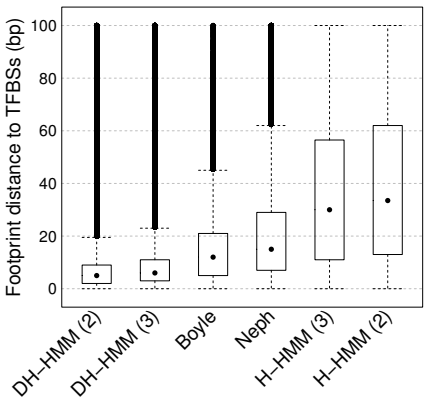

Fig. 3. Distribution of the distances from genomic positions predicted by segmentation-based methods to the center of the predicted active TFBSs. Results correspond to the absolute distances over all TFs using data from H1-hESC and K562 cell types.

5 DISCUSSION

Methods for detection of active TFBSs can be categorized in two main classes: site-centric (Cuellar and Centipede) and segmentation-based (Boyle, Neph and HMMs). Site-centric approaches require the identification of all MPBSs of a given TF to classify them as active or inactive. One advantage of the latter methods is that they have an ideal spatial specificity. On the other hand, they require sequence binding affinity information (PWM) of the TFs to be known *a priori*. On the other hand, segmentation-based approaches can be used for *de novo* motif detection. Indeed, previous studies have shown that footprints allowed the automatic creation of high-sensitivity TF models (Kulakovskiy, I. V. *et al.*, 2009) and the discovery of hundreds of novel motifs (Neph, S. *et al.*, 2012). Such studies would benefit from segmentation-based methods with good spatial specificity. Moreover, segmentation-based methods tend to reduce computational complexity, as they decrease drastically (1-2%) the genomic space used for motif detection. However, it is important to mention that the interpretation of such reduced genomic space would still require knowledge of proteins' binding sequence affinity.

While Centipede obtained good AUC results, the method's performance was too dependent on regularization parameters. Moreover, for large input files (TFs with a large number of PWM hits) the method required up to 6/core days of computing and 65GB of memory on a single TF and cellular condition. The most computational expensive segmentation method, DH-HMM, requires 9 days/core for predicting footprints and active binding sites over all 500 TFs from JASPAR in one cell type.

Another important aspect is the proposed combined use of DHS and histone modifications shapes around the TFBSs. Cuellar method is based on obtaining read counts around the TFBSs. Clearly, such method can not take the local shape profiles into account. For example, it can not distinguish the DHS peaks with the footprint signals and would detect any TFBS inside a region with high DHS levels as the regions indicated in Fig. 1A. Indeed, our experimental results confirm the poor specificity of the method. While Centipede used the local profiles of the DHS signal, it used simple read count statistics for the histone modification signals. Therefore, it is unable

to detect the valley shapes indicated in Fig. 1A. Not surprisingly, no improvements were possible with the use of histone modifications using Centipede, as indicated in Pique-Regi, R. *et al.* (2011). The DH-HMM model could indeed show improved results by using the local profiles of DHS and histone modification signals. Moreover, DH-HMM footprints cover a higher percentage of DHS regions than other segmentation-based methods.

Recent studies have also indicated that the histone modification profiles centered on DHS sites can be clustered and these clusters can also have asymmetric shapes, where there is lower evidence of histone marks downstream or upstream of the DHS. We have also tested variants of the DH-HMM to capture such asymmetric signals, but no significant improvement was obtained. As show in Supplementary Fig. 5, the slope signals have high values changes even on low histone modification values and are responsible for the correct characterization of such asymmetric peaks. Lastly, the HMMs displayed robustness in the training/evaluation on distinct cell types. This indicates that no further training of the HMMs and time intensive manual annotation of genomic regions are required. This is achieved by addressing both within- and between-dataset variability with our normalization pipeline.

6 FINAL REMARKS

This paper presents a novel approach to combine the spatial profiles of DHS (DNase-seq) and histone modifications (ChIP-seq) to predict cell type-specific active TFBSs. Moreover, we perform a large evaluation of all competing methods over two cell types and a large validation with 83 TF ChIP-seq datasets. We could show that the HMM model combining both DHS and histone modification data presents a good trade-off between sensitivity and specificity in relation to all compared methods. Furthermore, the method was robust when trained and evaluated in distinct cell types and has no further parameterization requirements. This study also provides all footprint predictions and validation data forming the first benchmarking data set for footprinting analyses. The accumulation of further epigenetic data and more detailed biological experiments will pose new methodological challenges to the field. The analysis of a cell type after a differentiation steps or response stimuli would indicate detailed changes in transcriptional landscape.

ACKNOWLEDGEMENTS

We would like to thank Pablo A. Jaskowiak, Sonja Haenzelmann, Manuel Allhoff, Joseph Kuo, Terry Furey and Shane Neph for providing predictions and sharing code and the anonymous referees for relevant suggestions.

Funding: This work was supported by the Interdisciplinary Center for Clinical Research (IZKF Aachen), RWTH Aachen University Medical School, Aachen, Germany; and Brazilian research agencies: FACEPE and CNPq.

REFERENCES

Arvey, A. *et al.* (2012). Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome Research*, **22**(9), 1723–1734.

Bell, O. *et al.* (2011). Determinants and dynamics of genome accessibility. *Nat Rev Genet*, **12**(8), 554–564.

Boyle, A. P. *et al.* (2008). F-seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics*, **24**(21), 2537–2538.

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.

Cock, P. J. A. *et al.* (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**(11), 1422–1423.

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.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**(7414), 57–74.

Gusmão, E. G. *et al.* (2012). Prediction of transcription factor binding sites by integrating dnase digestion and histone modification. In *Proc. of the 7th Brazilian Symposium on Bioinformatics*, Campo Grande, Mato Grosso do Sul, Brazil.

Hon, G. *et al.* (2009). Discovery and Annotation of Functional Chromatin Signatures in the Human Genome. *PLoS Comput Biol*, **5**(11), e1000566+.

Kim, J. *et al.* (2008). An extended transcriptional network for pluripotency of embryonic stem cells. *Cell*, **132**(6), 1049–1061.

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.

Landt, S. G. *et al.* (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Research*, **22**(9), 1813–1831.

Madden, H. H. (1978). Comments on the Savitzky-Golay convolution method for least-squares fit smoothing and differentiation of digital data. *Anal.Chem.*, **50**, 1383–1386.

Maston, G. A. *et al.* (2006). Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics*, **7**(1), 29–59.

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 TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic acids research*, **34**(Database issue), D108–D110.

Natarajan, A. *et al.* (2012). Predicting cell-type-specific gene expression from regions of open chromatin. *Genome Research*, **22**(9), 1711–1722.

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

Ouyang, Z. *et al.* (2009). ChIP-seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells. *Proceedings of the National Academy of Sciences*, **106**(51), 21521–21526.

Pedregosa, F. *et al.* (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

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.

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).

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, **16**(1), 16–23.

Thurman, R. E. *et al.* (2012). The accessible chromatin landscape of the human genome. *Nature*, **489**(7414), 75–82.

Wang, J. *et al.* (2013). Factorbook.org: a wiki-based database for transcription factor-binding data generated by the ENCODE consortium. *Nucleic Acids Research*, **41**(D1), D171–D176.

Whitington, 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.

Wilbanks, E. G. and Facciotti, M. T. (2010). Evaluation of algorithm performance in ChIP-seq peak detection. *PloS one*, **5**(7), e11471+.

Wilczynski, B. *et al.* (2009). Finding evolutionarily conserved cis-regulatory modules with a universal set of motifs. *BMC bioinformatics*, **10**(1), 82+.

Won, K. J. *et al.* (2010). Genome-wide prediction of transcription factor binding sites using an integrated model. *Genome Biology*, **11**(1), R7+.

Zhang, Y. *et al.* (2008). Model-based analysis of ChIP-Seq (MACS). *Genome biology*, **9**(9), R137+.