BIOINFORMATICS DISCOVERY NOTE Vol. 26 no. 17 2010, pages 2071–2075
doi:10.1093/bioinformatics/btq405

Genome analysis

Advance Access publication July 27, 2010

Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites

Stephen A. Ramsey, Theo A. Knijnenburg, Kathleen A. Kennedy, Daniel E. Zak, Mark Gilchrist, Elizabeth S. Gold, Carrie D. Johnson, Aaron E. Lampano, Vladimir Litvak, Garnet Navarro, Tetyana Stolyar, Alan Aderem* and Ilya Shmulevich*

Institute for Systems Biology, 1441 North 34th Street, Seattle, WA, 98103, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Histone acetylation (HAc) is associated with open chromatin, and HAc has been shown to facilitate transcription factor (TF) binding in mammalian cells. In the innate immune system context, epigenetic studies strongly implicate HAc in the transcriptional response of activated macrophages. We hypothesized that using data from large-scale sequencing of a HAc chromatin immunoprecipitation assay (ChIP-Seq) would improve the performance of computational prediction of binding locations of TFs mediating the response to a signaling event, namely, macrophage activation.

Results: We tested this hypothesis using a multi-evidence approach for predicting binding sites. As a training/test dataset, we used ChIP-Seq-derived TF binding site locations for five TFs in activated murine macrophages. Our model combined TF binding site motif scanning with evidence from sequence-based sources and from HAc ChIP-Seq data, using a weighted sum of thresholded scores. We find that using HAc data significantly improves the performance of motif-based TF binding site prediction. Furthermore, we find that within regions of high HAc, local minima of the HAc ChIP-Seq signal are particularly strongly correlated with TF binding locations. Our model, using motif scanning and HAc local minima, improves the sensitivity for TF binding site prediction by ~50% over a model based on motif scanning alone, at a false positive rate cutoff of 0.01.

Availability: The data and software source code for model training and validation are freely available online at http://magnet.systemsbiology.net/hac.

Contact: aderem@systemsbiology.org; ishmulevich@systemsbiology.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 27, 2010; revised on June 30, 2010; accepted on July 3, 2010

1 INTRODUCTION

Mammalian cells exhibit diverse transcriptional profiles across different cell types and conditions, for example, in immune cells activated with different pathogen-associated molecules (Ramsey *et al.*, 2008). To a large extent, these profiles are controlled by the arrangement and chromatin accessibility of *cis*-regulatory elements (Berger, 2007). Transcription factors (TFs) bind specific sequence elements in chromatin locations of permissive epigenetic or conformational states, leading to activation or repression of transcriptional activity. For mapping these regulatory interactions, it is particularly promising that the binding of a TF can now be measured genome wide using chromatin immunoprecipitation (IP) with sequence detection (ChIP-Seq, see Johnson *et al.*, 2007). However, antibody and cellular material requirements preclude using ChIP-Seq to screen for all TFs mediating a transcriptional response. There remains a need for computational approaches that can, in the absence of experimental TF binding data, leverage transcriptional data and genomic information to identify the network of TFs and binding sites that underlies a transcriptional response.

An important tool for predicting mammalian TF binding sites is motif scanning, i.e. searching DNA sequence for matches within a library of sequence motifs reported to be bound by specific TFs (Lähdesmäki *et al.*, 2008). Such a library enables mapping between a scanning-identified sequence element and one or more candidate TFs that may bind it. However, such motifs are often highly uncertain and they can be degenerate, leading to a high frequency of false positive predictions (Hannenhalli, 2008). Furthermore, mammalian *cis*-regulatory elements can be tens of kilobases from transcription start sites, necessitating searching large sequence regions and further increasing false positives. These issues undermine the performance of motif scanning as a standalone approach. Successful motif-based prediction of TF binding depends on identifying the sequence regions, within the relevant cell type, that are likely to contain *cis*-regulatory elements (Ernst *et al.*, 2010; Wasserman and Sandelin, 2004; Whitington *et al.*, 2009).

It has been observed that *cis*-regulatory elements tend to co-occur with chromatin or sequence features that can be grouped in three categories: (i) chromatin structural features such as DNase I hypersensitive sites; (ii) epigenetic marks such as histone acetylation (HAc); and (iii) sequence features such as high GC content and conservation across species. The HAc mark, which has been associated with active promoters and open chromatin (Vettese-Dadey *et al.*, 1996), is of particular relevance to transcriptional regulation because the modification can be placed or removed in response to the cellular state. These observations have spurred the development of approaches that integrate data for multiple types of chromatin features to improve the accuracy of TF binding site predictions. Various data integration frameworks for binding site prediction have been used, including the support vector machine (Holloway *et al.*, 2005; Nykter *et al.*, 2009), probabilistic methods

*To whom correspondence should be addressed.

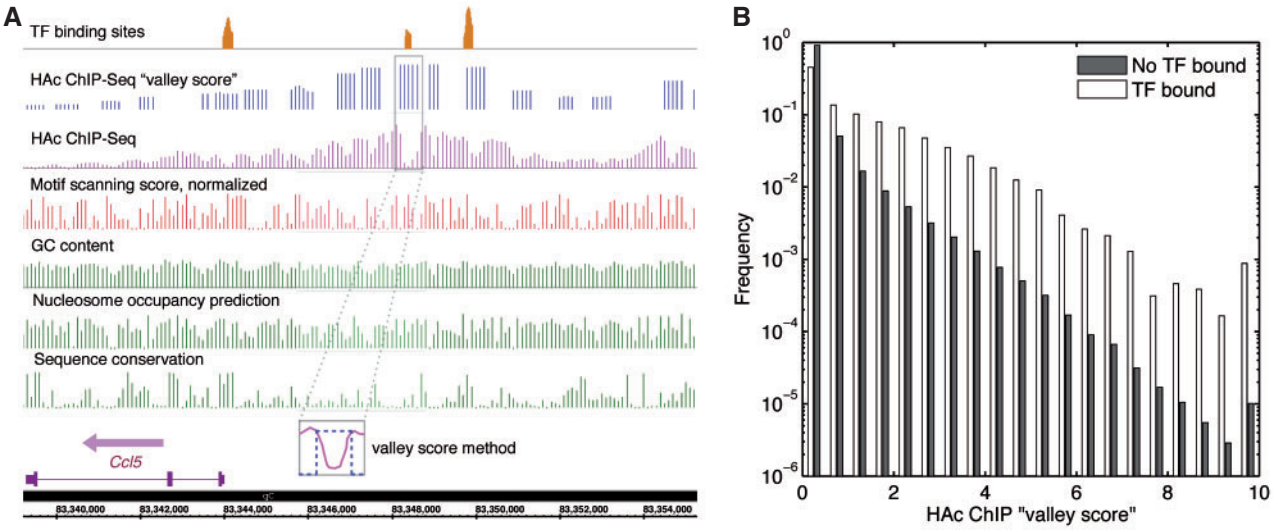

Fig. 1. Local minima in the HAc ChIP signal correlate with TF binding. (**A**) A 12 kbp region of mouse chromosome 11 including the gene *Ccl5* [an LPS-regulated cytokine with multiple NFκB (nuclear factor of kappa light polypeptide gene enhancer in B-cells) sites in its promoter] and its upstream regulatory region, with TF binding data and various feature tracks. Measured binding locations for the TF NFκB/p50 are shown in orange (top row). Each feature track is displayed in colored vertical bars interspersed every 100 bp: HAc ChIP-Seq signal (magenta); 'valley scores' (VS) for HAc local minima (blue); normalized NFκB binding site motif match scores (red); DNA sequence GC content (green); nucleosome occupancy score (green); and vertebrate conservation score for genomic sequence (green). The NFκB binding sites correspond to local minima in the HAc ChIP-Seq signal. Inset: Within a local minimum of the ChIP-Seq signal (magenta curve), the smaller of the maximum signal values on either side of the local minimum is computed, and the entire local minimum region is assigned that value as its VS, and the value zero outside the local minimum region (blue lines) (Supplementary Section S1.6). (**B**) The distribution of HAc ChIP VS from stimulated cells in TF-bound sites differs substantially from non-TF-bound sites, as shown in the two histograms (note the logarithmic vertical scale).

(Beyer *et al.*, 2006; Ernst *et al.*, 2010; Lähdesmäki *et al.*, 2008), and a kernel-based classifier (Wang *et al.*, 2009). Early studies integrating genomic data into binding site prediction were carried out in yeast (Beyer *et al.*, 2006; Holloway *et al.*, 2005), or in mammals using ground-truth datasets that were not cell type-specific (Lähdesmäki *et al.*, 2008). More recent approaches have used cell type-specific mammalian epigenetic or transcriptional data to predict binding for a single TF in mammals (Nykter *et al.*, 2009; Wang *et al.*, 2009). Other recent studies have used genome-wide datasets for multiple TFs to develop prediction models without directly incorporating epigenetic data into the model (Won *et al.*, 2009; Zhou *et al.*, 2010). Two recent studies incorporated histone methylation ChIP-Seq data into multi-evidence prediction models, using ChIP-derived ground-truth datasets of 10 and 13 TFs, respectively (Whitington *et al.*, 2009; Won *et al.*, 2010). Whitington *et al.* found that predictions are improved when the methylation data are derived from the same tissue type from which the TF binding site measurements are derived.

In this study, we investigated the hypothesis that incorporating HAc ChIP-Seq data into a multi-evidence, motif scanning-based model can improve TF binding site predictions. We further studied whether prediction performance is improved when the HAc data are derived from the same cell condition from which the TF binding data (used for evaluating performance) are derived. Having observed that TF binding locations frequently occur at local minima of HAc ChIP-Seq signal within regions of high HAc (see Fig. 1 and Supplementary Fig. S1), we also studied the predictive utility of 'valley scores' (VS) assigned to local minima of the HAc ChIP-Seq signal. Following our previous investigation of the regulatory network underlying macrophage activation (Ramsey *et al.*, 2008), this study was carried out using TF binding and HAc measurements in the macrophage, a key cell type of the innate immune system. When activated by exposure to a pathogen-associated molecule such as lipopolysaccharide (LPS), the macrophage undergoes extensive transcriptional reprogramming that is mediated in part by alterations in HAc (Aung *et al.*, 2006).

2 APPROACH

The HAc hypothesis was tested using an integrative TF binding site (TFBS) prediction framework and using an approach designed to estimate the performance that the prediction model would have on a novel TF for which only a binding site motif is available. As features, the framework used motif scanning data (Supplementary Fig. S2) along with subsets of seven non-TF-specific features selected for their potential association with TFBSs. As shown in Figure 1A and Supplementary Table S1, the features consisted of HAc (acetylated H4) ChIP-Seq data from activated and non-activated macrophages; VS derived from the HAc data; and three features based on genomic sequence (GC content, vertebrate species conservation and a nucleosome occupancy prediction score). A peak in the HAc VS signal corresponds to a local minimum in the HAc ChIP-Seq signal. As a ground-truth TFBS dataset, we used ChIP-Seq data, from activated macrophages, for five TFs (Supplementary Table S2). In keeping with the study goals, the ChIP-Seq data were *not* used to improve the motifs, and model performance was tested using binding data for a TF that was not used in the model training. Performance measurements obtained using such a TF-based cross-validation are, in our view, more relevant to this application (library-based motif

scanning) than are results from chromosome-based cross-validation. Importantly, HAc was measured in LPS-stimulated macrophages, consistent with the conditions for TF binding measurements.

TF binding site predictions were made in adjacent 100 bp intervals spanning 10 kb promoter regions of genes that are expressed in murine macrophages. A value for each TF prediction feature was computed within each 100 bp interval, from the feature's raw data. A weighted, thresholded linear model class was used to combine the motif scanning feature with zero, one or two additional features to predict binding for the five TFs. This model class divides the range of each feature's values into three regimes: below minimum threshold (value changes within this regime are not informative), above the maximum threshold (saturated; changes are also not informative) and within the linear response range. Fifteen models, each using a different combination of features (Supplementary Table S3), were trained using the ground-truth dataset. Because the model using HAc ChIP VS performed the best among the two-feature models, the three-feature model analysis was restricted to models with motifs, HAc ChIP VS and one additional feature. For training, each model's parameters were optimized to maximize the average prediction performance for a set of four TFs, with the performance metric being the area under the sensitivity versus false positive rate (FPR) curve. The performance of the model, with the best parameter set from the training, was then tested on the fifth TF, and averaged over the leave-one-out cross-validation.

3 METHODS

Complete methods are described in Supplementary Material, Section S1.

Ground-truth dataset: ChIP-Seq assays were performed for the TFs ATF3, C/EBPδ, IRF1, NFκB/p50 and NFκB/p65 in macrophages activated through treatment with purified Toll-like receptor agonists for 1–6 h (see Supplementary Table S2 and Section S1.4). Binding locations were identified from above-threshold locations in the ChIP-Seq signal, as described in Supplementary Section S1.7.

Prediction features: TF predictions were made in 100 bp intervals (as used in Won *et al.*, 2010) of transcript-proximal regions comprising ∼7% of the genome, selected as described in Supplementary Section S1.2. Combinations of eight features, individually listed in Supplementary Table S1 and labeled by index f, were used for TF binding prediction. Feature $f=1$, which conferred TF specificity to the predictions, was based on motif scanning. For each TF, motif position-weight matrices (PWMs) corresponding to the TF were obtained from TRANSFAC (Supplementary Table S2 and Section S1.3). Sequences were scanned for motif matches using a likelihood-based algorithm (Lähdesmäki *et al.*, 2008), and combined to obtain, within each interval and for each TF, a score representing the strength of the best match for any motif corresponding to that TF, at any position within the interval. Features 2–5 of Supplementary Table S1 were derived from HAc ChIP-Seq assays of unstimulated macrophages or macrophages stimulated for 1, 4 or 6 h with LPS (Supplementary Sections S1.4–1.5). VS for HAc local minima were computed as described in Supplementary Section S1.6. Features 6–8 were based on genomic sequence, and thus are not macrophage specific. For the stimulated-cell HAc ChIP-Seq features (Supplementary Table S1, rows 2 and 4), the time point for the HAc dataset that was used was always the same as the time point of the ground-truth dataset for the TF for which predictions were being made.

Prediction model: within each interval i, the model integrates a set F of up to three features (always including the motif feature, $f=1$) by a weighted sum of thresholded feature values. Feature values may depend on the TF t, as is the case for motif scanning, or on the cellular condition for which TF binding predictions are being made (as is the case for HAc-derived features). The value for feature f at interval i and TF t is therefore denoted by v_{fit}. The feature value v_{fit} is passed through a piecewise-linear function θ_f that is defined by feature-specific thresholds λ_f and μ_f,

$$\theta_f(v)=\begin{cases}0, & v<\lambda_f,\\ (v-\lambda_f)/(\mu_f-\lambda_f), & \lambda_f\le v\le\mu_f,\\ 1, & v>\mu_f.\end{cases} \quad (1)$$

The prediction score σ_{it} that the TF t binds within interval i is obtained by a weighted sum of thresholded contributions, but with a multiplicative factor enforcing a minimum TF-specific motif match value for a non-zero σ_{it},

$$\sigma_{it}=\theta(\theta_1(v_{1it}))\left(\sum_{f\in F}\omega_f\theta_f(v_{fit})\right), \quad (2)$$

where the weight vector $\vec{\omega}$ has unit L1 norm (a negative component would represent a feature that is anti-correlated with TF binding), and where θ is defined by $\theta(x)=0$ if $x\le 0$ and $\theta(x)=1$ if $x>0$. Importantly, a given model instance $\mathcal{M}$, defined by the tuple $\{F,\vec{\lambda},\vec{\mu},\vec{\omega}\}$, is TF independent.

Performance metric: for a given model $\mathcal{M}$, TF t, and prediction score cutoff σ, the set of intervals $\Pi(\sigma,t)$ for which $\sigma_{it}\ge\sigma$ were predicted to contain binding sites for t (remaining intervals were predicted to have no t binding). The set of intervals containing ground-truth binding sites (based on ChIP-Seq) is denoted by $\Sigma(t)$. Because the typical ChIP-Seq fragment size was ∼160 bp, some TF binding locations appeared as adjacent intervals in $\Sigma(t)$; these were counted as single binding sites. The number of ground-truth binding sites $B(t)$ was counted (Supplementary Table S2), and the fraction of these binding sites that coincided with at least one interval $i\in\Pi(\sigma,t)$, was computed as the sensitivity $S(\sigma,t)$. The FPR $E(\sigma,t)$ was computed by dividing the number of intervals in the set difference $\Pi(\sigma,t)\setminus\Sigma(t)$ by the number of intervals not contained in $\Sigma(t)$. The cutoff σ was varied and the resulting $(E(\sigma,t),S(\sigma,t))$ function [receiver operating characteristic (ROC) curve] was numerically integrated over the range $0<E\le 0.01$ to obtain the TF-specific performance score $A(t)$. For model training (Supplementary Section S1.12), the cost function used was $C(t)=1-A(t)/0.01$. During training, cases where it was not possible to obtain a sufficient number of (S,E) samples were handled using a penalty, as described in Supplementary Section S1.11.

Model training: groups of four TFs at a time were selected for model training, and for a given model $\mathcal{M}$, the cost was averaged over the four TFs, $C=\langle C(t)\rangle_t$. Model parameters were varied to minimize C subject to constraints on $\vec{\lambda}$, $\vec{\mu}$ and $\vec{\omega}$, using a two-stage optimization process (Supplementary Section S1.12), to obtain the best parameter set for the model with features F.

Model testing: for both training and testing purposes, the performance $A(t')$ of the model with the best parameter set from the training, was measured on the fifth TF t' using leave-one-out cross-validation. The five values for $A(t')$ were compared between different feature groups F using a paired t-test, and summarized in terms of the mean and SD (Supplementary Table S3).

4 RESULTS

Feature distributions: first, the TF specificity of the motif scanning was investigated. Across all five TFs, the motif scanning score distribution from TFBSs was significantly higher than the distribution from non-binding sites (Supplementary Fig. S2). Next, HAc VS representing local minima were computed, and the distributions of VS at TFBSs and non-binding sites were compared. In LPS-stimulated cells, HAc VS were significantly higher at TFBSs than at non-binding sites (Fig. 1B); this motivated the use of HAc ChIP data to improve predictions. Furthermore, LPS-dependent TFBSs were correlated with LPS-inducible HAc local minima (Supplementary Table S4 and Fig. S3).

Model performance: first, two-feature models (motifs plus one other feature) were compared with a motifs-only reference model. Based on the area under the sensitivity versus FPR curve (Fig. 2, Supplementary Fig. S4 and Table S3), the model with HAc VS from

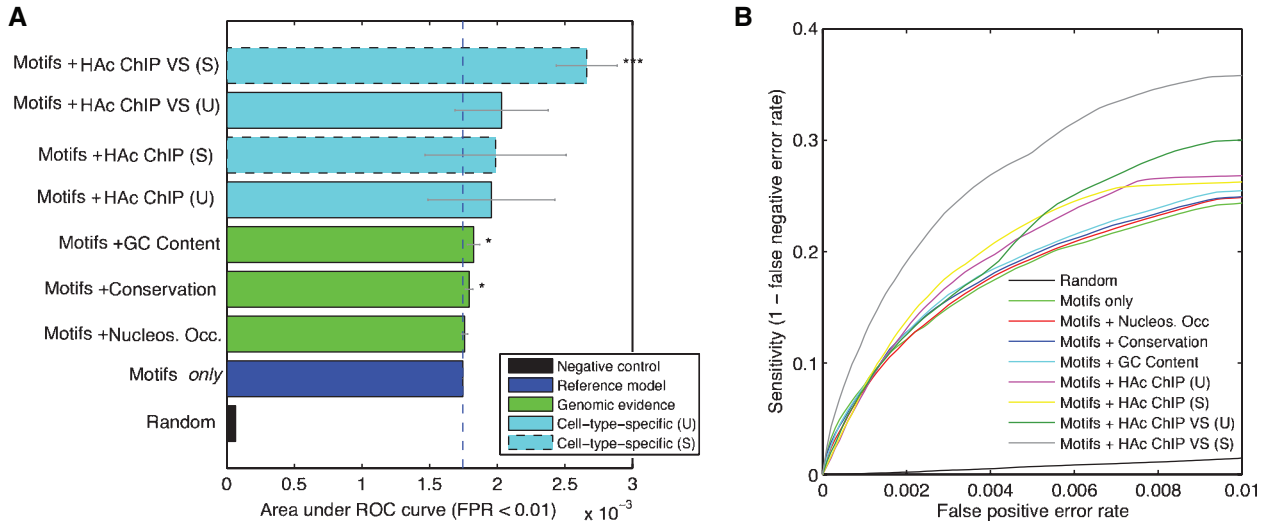

Fig. 2. HAc data improve motif scanning-based TFBS predictions. (**A**) Prediction performance (area under the sensitivity versus FPR curve, or 'ROC' curve) for models with motif scanning and one additional feature, and a motifs-only reference model (data for models with three features are shown in Supplementary Fig. S4). Larger bar values correspond to better cross-validation-average performance on the test dataset. The performance for the reference model is shown in the blue bar (and vertical dotted line), and a random model is shown as a negative control (black bar). The motifs-only model outperformed the random model, ~27-fold. Each green bar represents a model that used motif information plus a specific sequence-based feature (GC content, etc.). Each cyan bar represents a model that used motif information plus a HAc ChIP-Seq-based feature (Supplementary Table S3). Each error bar represents the cross-validation-wide SD of the performance difference between the indicated model and the reference model (Section 3). $^{*}P < 0.05$; $^{***}P < 0.001$. For the cyan bars, a dashed border indicates that HAc data are from LPS-stimulated cells; a solid border means the HAc data were from unstimulated cells. In the top two bar labels, 'VS' stands for the 'valley score' for local minima in the HAc ChIP-Seq signal. (**B**) ROC curves, for predictions by the models shown in (A) (see Supplementary Fig. S4 for the complete FPR range). The model with HAc VS (from stimulated cells; gray curve) outperforms the other models. ROC curves were obtained by varying the prediction score cutoff (Section 3). The lack of improvement for the nucleosome occupancy-based model is consistent with the very weak association between this feature and TF binding (Supplementary Fig. S6).

stimulated cells had the highest performance improvement relative to the reference model (52% increase, $P < 10^{-3}$). The HAc ChIP-Seq signal also improved prediction performance (by 14%), but the improvement was highly variable from TF to TF (coefficient of variation = 27%; see Supplementary Table S5). The model using the stimulated-cell HAc VS also outperformed the unstimulated-cell HAc VS data (by 31%, $P < 0.01$). In contrast to the HAc ChIP-derived datasets, the three genomic features (GC content, conservation and nucleosome occupancy score) did not substantially improve prediction performance. However, the improvements due to GC content (5% increase) and conservation (3%) were more consistent from TF to TF, and thus in both cases were statistically significant ($P < 0.05$). Next, models with motifs plus two other features were compared with the best previous model (motifs + HAc VS). None of the models gave a statistically significant improvement over the best two-feature model (Supplementary Fig. S5). These findings suggest that more TF binding data would be required to discriminate prediction performances of three-feature models.

5 CONCLUSIONS

Using cell type-specific HAc ChIP-Seq data improves motif scanning-based prediction of TFBSs in primary macrophages. This prediction strategy could be applied to any cell type in which HAc can be globally measured. Overall, these findings suggest that within histone-acetylated regions, local minima of HAc ChIP-Seq signal may indicate sites of active transcriptional regulation.

ACKNOWLEDGEMENTS

We thank K. Deutsch, S. Bloom and M. Gundapuneni for technical assistance. H. Lähdesmäki and M. Nykter kindly provided some MATLAB functions. S.A.R. thanks A. Diercks, E. Fu and V. Thorsson for helpful discussions. We thank A. Nachman, B. Marzolf, D. Rodriguez and L. Rowen for coordinating the contributions of their groups.

Funding: The National Heart, Lung, and Blood Institute (K25HL098807 to S.A.R.); the National Institute of Allergy and Infectious Diseases (HHSN272200700038C); and the National Institute of General Medical Sciences (R01GM072855 to I.S. and P50GM076547).

Conflict of Interest: none declared.

REFERENCES

Aung,H.T. *et al.* (2006) LPS regulates proinflammatory gene expression in macrophages by altering histone deacetylase expression. *FASEB J.*, **20**, 1315–1327.

Berger,S.L. (2007) The complex language of chromatin regulation during transcription. *Nature*, **447**, 407–412.

Beyer,A. *et al.* (2006) Integrated assessment and prediction of transcription factor binding. *PLoS Comput. Biol.*, **2**, e70.

Ernst,J. *et al.* (2010) Integrating multiple evidence sources to predict transcription factor binding in the human genome. *Genome Res.*, **20**, 526–536.

Hannenhalli,S. (2008) Eukaryotic transcription factor binding sites–modeling and integrative search methods. *Bioinformatics*, **24**, 1325–1331.

Holloway,D.T. *et al.* (2005) Integrating genomic data to predict transcription factor binding. *Genome Inform.*, **16**, 83–94.

Johnson,D.S. *et al*. (2007) Genome-wide mapping of *in vivo* protein-DNA interactions. *Science*, **316**, 1497–1502.

Lähdesmäki,H. *et al*. (2008) Probabilistic inference of transcription factor binding from multiple data sources. *PLoS ONE*, **3**, e1820.

Nykter,M. *et al*. (2009) A data integration framework for prediction of transcription factor targets: a BCL6 case study. *Ann. N. Y. Acad. Sci.*, **1158**, 205–214.

Ramsey,S.A. *et al.* (2008) Uncovering a macrophage transcriptional program by integrating evidence from motif scanning and expression dynamics. *PLoS Comput. Biol.*, **4**, e1000021.

Vettese-Dadey,M. *et al.* (1996) Acetylation of histone H4 plays a primary role in enhancing transcription factor binding to nucleosomal DNA *in vitro*. *EMBO J.*, **15**, 2508–2518.

Wang,T. *et al.* (2009) A general integrative genomic feature transcription factor binding site prediction method applied to analysis of USF1 binding in cardiovascular disease. *Hum. Genomics*, **3**, 221–235.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.

Whitington,T. *et al.* (2009) High-throughput chromatin information enables accurate tissue-specific prediction of transcription factor binding sites. *Nucleic Acids Res.*, **37**, 14–25.

Won,K.-J. *et al.* (2009) An integrated approach to identifying *cis*-regulatory modules in the human genome. *PLoS One*, **4**, e5501.

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

Zhou,X. *et al.* (2010) A systems biology approach to transcription factor binding site prediction. *PLoS One*, **5**, e9878.