BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 1 2012, pages 56–62
doi:10.1093/bioinformatics/btr614

Gene expression

Advance Access publication November 8, 2011

Epigenetic priors for identifying active transcription factor binding sites

Gabriel Cuellar-Partida[1], Fabian A. Buske[1], Robert C. McLeay[1], Tom Whitington[1], William Stafford Noble[2,3] and Timothy L. Bailey[1,*]

[1]Institute for Molecular Bioscience, The University of Queensland, Brisbane QLD 4072, Australia, [2]Department of Genome Sciences and [3]Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA

Associate Editor: Janet Kelso

ABSTRACT

Motivation Accurate knowledge of the genome-wide binding of transcription factors in a particular cell type or under a particular condition is necessary for understanding transcriptional regulation. Using epigenetic data such as histone modification and DNase I, accessibility data has been shown to improve motif-based *in silico* methods for predicting such binding, but this approach has not yet been fully explored.

Results We describe a probabilistic method for combining one or more tracks of epigenetic data with a standard DNA sequence motif model to improve our ability to identify active transcription factor binding sites (TFBSs). We convert each data type into a position-specific probabilistic prior and combine these priors with a traditional probabilistic motif model to compute a log-posterior odds score. Our experiments, using histone modifications H3K4me1, H3K4me3, H3K9ac and H3K27ac, as well as DNase I sensitivity, show conclusively that the log-posterior odds score consistently outperforms a simple binary filter based on the same data. We also show that our approach performs competitively with a more complex method, CENTIPEDE, and suggest that the relative simplicity of the log-posterior odds scoring method makes it an appealing and very general method for identifying functional TFBSs on the basis of DNA and epigenetic evidence.

Availability and implementation: FIMO, part of the MEME Suite software toolkit, now supports log-posterior odds scoring using position-specific priors for motif search. A web server and source code are available at http://meme.nbcr.net. Utilities for creating priors are at http://research.imb.uq.edu.au/t.bailey/SD/Cuellar2011.

Contact: t.bailey@uq.edu.au

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

Received on July 22, 2011; revised on October 26, 2011; accepted on October 31, 2011

1 INTRODUCTION

Binding of transcription factors (TFs) to DNA is a key step in the regulation of gene expression, strongly influencing the activation or repression of genes. However, accurately locating transcription factor binding sites (TFBSs) in eukaryotic genomes containing gigabases of DNA is challenging because a typical TF binds to relatively short (4–10 bp) DNA sequences that occur at many places throughout the genome, only a small portion of which are actively involved in gene regulation in any particular cell type.

In vivo, several molecular mechanisms limit the binding of TFs to a small fraction of the potential sites. Perhaps most significantly, local chromatin structure, arising from the packaging of DNA, histones and other proteins, occludes access of the TF to many genomic loci. Some aspects of this structure—notably, numerous types of post-translational covalent modifications such as acetylation, methylation, phosphorylation, ubiquitination, ribosylation, glycosylation and sumoylation—affect TF binding in a complex fashion that is not yet completely understood. Empirically, however, numerous studies have shown that TF binding is associated with several types of histone marks (e.g mono- and tri-methylation of histone H3 lysine 4) (Barski *et al.*, 2007; Cui *et al.*, 2009; Heintzman *et al.*, 2007; Heintzmann *et al.*, 2009; Robertson *et al.*, 2007) and with hypersensitivity to cleavage by DNase I (Bernat *et al.*, 2006; Crawford *et al.*, 2006; Hesselberth *et al.*, 2009; Keene *et al.*, 1981; McArthur *et al.*, 2001; Wu, 1980).

In light of these observations, a variety of computational methods have been proposed that aim to improve our ability to identify active TFBSs on the basis of the DNA sequence plus one or more epigenetic experimental assays. Perhaps the simplest such approach scans for occurrences of a specified DNA motif, but deterministically filters out some positions on the basis of the raw signal from a histone modification ChIP-seq assay (Whitington *et al.*, 2009). This type of filter greatly improves the accuracy of TFBS prediction. Several other integrative methods have been proposed for combining DNA motifs with chromatin data (Ernst *et al.*, 2010; Won *et al.*, 2009, 2010). Among these methods, the one most directly comparable to the one proposed here is CENTIPEDE (Pique-Regi *et al.*, 2011), which employs a hierarchical Bayesian mixture model that incorporates information about the DNA sequence, evolutionary conservation, distance to the transcription start site, DNase I accessibility and activating and repressing histone marks.

In this work, we propose a probabilistic method for combining one more more tracks of epigenetic data with a standard DNA sequence motif model to improve our ability to identify active TFBSs. Our work thus makes two primary methodological contributions. First, we describe a heuristic procedure for converting each data type into a probabilistic prior. The approach itself is novel, though it is motivated by similar methods introduced in the context of

*To whom correspondence should be addressed.

motif discovery by the Hartemink group (Gordân *et al.*, 2010; Narlikar *et al.*, 2006), and in the context of TFBS prediction by Lahdesmaki *et al.* (2008). Critically, our method is applicable to any real-valued data type, requiring the specification of a single hyperparameter with an intuitive interpretation. Second, we describe how to combine a given sequence motif, represented as a position weight matrix (PWM), with the prior to compute a log-posterior odds score. This approach is directly analogous to the most widely used method for scanning for occurrences of TFBSs, and contrasts with a more computationally intensive Markov Chain Monte Carlo-based method proposed previously (Lahdesmaki *et al.*, 2008).

In addition to these methodological contributions, our work suggests three significant conclusions. First, our experiments, using histone modifications H3K4me1, H3K4me3, H3K9ac and H3K27ac, as well as DNase I sensitivity, show conclusively that the log-posterior odds score consistently outperforms a previously described binary filter based on the same data (Whitington *et al.*, 2009). This point is critical, because the simple filter has clear intuitive appeal; we need to begin by demonstrating that our method improves upon this simple baseline. Second, among epigenetic datasets, we find that DNase I sensitivity provides the most value, with only a relatively small additional boost in performance when we also consider histone marks. Third, we evaluate our method using the benchmark dataset described in the CENTIPEDE paper. In this setting, CENTIPEDE performs better on average, but the improvement is not uniform across TFs. This comparison to CENTIPEDE clearly illustrates the value of our proposed method. The strong performance of the log-posterior odds score, coupled with the relative simplicity of the method, make it an appealing and very general method for identifying functional TFBSs on the basis of DNA and epigenetic evidence. In practical terms, the second and third conclusions above combine to suggest that a single prior based on DNase I is all we need to achieve excellent TFBS identification in practice.

2 METHODS

2.1 Creating priors from epigenetic data

We desire a prior distribution on the probability that genomic position i is bound by any TF in a given tissue. (Throughout the article, we use the terms 'tissue' and 'tissue-specific' to refer to data from a particular tissue, cell line or condition.) If we have tissue- and position-specific information y_i for each genomic position, $1 \le i \le n$, we denote the desired prior distribution by $Pr(B_i|y_i)$, where the random variable B_i is true if and only if *any* TF is bound at position i. The y_i can be, for example, tag counts at genomic positions from a histone modification or DNase I assay. We use a heuristic procedure to convert these tag counts to a function that we can use as the desired prior distribution.

To carry out the conversion from tag count to prior, we require a monotonically increasing function, $f(y_i)$, such that $0 \le f(y_i) \le 1$ for all values of y_i. Although we use $f(y_i)$ as a prior distribution, we do not require that the n transformed y_i values sum to one. Rather, we would like them to sum to some user-chosen estimate, $\beta > 0$, of the total number of binding sites of *all* TFs in the given tissue.

To construct the mapping function $f(y_i)$, we first map the interval bounded by the minimum and maximum observed values of y_i ($[y_{\min}, y_{\max}]$), to the interval $[a, 1]$ using the linear function

$$g(y_i) = \frac{1-a}{y_{\max}-y_{\min}}(y_i - y_{\min}) + a,$$

where $0 < a < 1$. In a moment we will see that the value of a is not free, but is determined by our choice of β. In this work, we employ a linear mapping, but any monotonically increasing function could be used for $g(y_i)$.

We then define $f(y_i)$ in terms of $g(y_i)$ as

$$f(y_i) = \beta \frac{g(y_i)}{\sum_{j=1}^{n} g(y_j)}. \quad (1)$$

It is trivial to see that summing $f(y_i)$ over all genomic positions gives β, as desired. Since $g(y_i)$ is monotonic and increasing and β is positive, $f(y_i)$ is clearly monotonic and increasing, as required. We also require that $f(y_i) \ge 0$, and we note that this is true because $g(y_i)$ and β are both positive. To ensure that $f(y_i) \le 1$, we note that the smallest that $g(y_i)$ can be is a, and the largest it can be is 1, so

$$\beta \frac{g(y_i)}{\sum_{j=1}^{n} g(y_j)} \le \beta \frac{1}{1+a(n-1)}.$$

Setting the right-hand side above equal to 1 and solving for a gives

$$a = \frac{\beta - 1}{n-1}. \quad (2)$$

Using this value for a ensures that $f(y_i) \le 1$. Note that since it makes no sense to set $\beta > n$ (because β is the prior estimate of the total number of binding sites in the given tissue, which cannot exceed the number of genomic positions n), a will always be < 1, as required to define the mapping function $g(y_i)$ above.

2.2 Log-posterior odds motif scoring

2.2.1 Definition In the previous section, we described a function on epigenetic data, $f(y_i)$, which we will interpret as $Pr(B_i|y_i)$, the prior probability that genomic position i is occupied by any TF in the tissue of the epigenetic data. Thus, $\Pr(B_i)$ represents the 'general binding propensity' of genomic region x_i. In this section, we derive a scoring function for predicting the TFBSs of a *given TF* in a *given tissue* using a standard PWM TF motif, M, and the position-specific prior function, $f(y_i)$. This scoring function, which we call the log-posterior odds score, is simply the sum of the traditional log-likelihood ratio motif score, $S(x_i)$, and the log-prior odds score, $P(i)$,

$$\hat{S}(x_i) = S(x_i) + P(i). \quad (3)$$

The log-prior odds score is defined as

$$P(i) = \log\left(\frac{f(y_i)}{1-f(y_i)}\right). \quad (4)$$

The log-likelihood ratio motif score is

$$S(x_i) = \sum_{j=1}^{w} M_{b,j},$$

where w is the width of the motif and $b = x_{i,j}$ refers to the DNA base in the j-th position of site x_i.

2.2.2 Derivation The log-posterior odds score [Equation (3)] is motivated by Bayes decision theory, which states that the best score for discriminating sites from non-sites given an observed sequence and any additional information is the posterior odds (Duda *et al.*, 2001). We will show that the log-posterior odds score defined by Equation (3) is approximately equal to the log-posterior odds that a candidate site for the given TF starts at position i in the sequence.

$$\hat{S}(x_i) \approx \log\left(\frac{\Pr(s_i|x_i, y_i)}{\Pr(\overline{s_i}|x_i, y_i)}\right). \quad (5)$$

Here, the Boolean variable s_i is true if and only if x_i is a binding site for the *given TF* in a given tissue. Note that Boolean variable s_i differs from B_i, which is true when *any TF* binds to site x_i in the given tissue.

We make the simplifying assumption that the sequence data x_i and epigenetic data y_i are *conditionally independent* given the 'class' variable

s_i. In other words, we assume that once we know whether x_i is a site or not, knowing y_i adds no additional information about the particular sequence of bases comprising x_i, and vice versa. Under this assumption, we can write

$$\Pr(x_i, y_i|s_i) = \Pr(x_i|s_i)\Pr(y_i|s_i). \quad (6)$$

By applying Bayes' rule twice, and by exploiting Equation (6), we can rewrite the numerator of Equation (5), the probability of the class being 'true' (e.g. position i is a site), as

$$\begin{aligned}\Pr(s_i|x_i, y_i) &= \frac{\Pr(x_i, y_i|s_i)\Pr(s_i)}{\Pr(x_i, y_i)} \\ &= \frac{\Pr(x_i|s_i)\Pr(y_i|s_i)\Pr(s_i)}{\Pr(x_i, y_i)} \\ &= \frac{\Pr(x_i|s_i)\Pr(s_i|y_i)\Pr(y_i)}{\Pr(x_i, y_i)}. \quad (7)\end{aligned}$$

A similar derivation rewrites the denominator of Equation (5), the probability of position i *not* being a site, as

$$\Pr(\overline{s_i}|x_i, y_i) = \frac{\Pr(x_i|\overline{s_i})\Pr(\overline{s_i}|y_i)\Pr(y_i)}{\Pr(x_i, y_i)}. \quad (8)$$

Thus, under the independence assumption above, the log-posterior ratio [(the right-hand side of Equation (5)] can be expressed as the sum of a log-likelihood ratio and the log-prior odds

$$\log\left(\frac{\Pr(x_i|s_i)}{\Pr(x_i|\overline{s_i})}\right) + \log\left(\frac{\Pr(s_i|y_i)}{\Pr(\overline{s_i}|y_i)}\right), \quad (9)$$

by applying Equations (7) and (8) and a bit of algebra. This approach can be easily extended to multiple priors under similar independence assumptions, adding additional log-prior odds terms to the above sum (Section 1 in Supplementary Material).

We estimate the log-likelihood ratio [the first term in Equation (9)] using the standard PWM score, $S(x_i)$. We assume that the binding sites of the given TF are correctly modeled by the position-specific probability matrix $\Theta_m = \{f_{c,j}\}$, where $f_{c,j}$ is the probability of DNA base c at position j in the motif of width w. We further assume that non-sites are modeled by a zero-order Markov model $\Theta_0 = \{p_c\}$, where p_c is the probability of base c at any non-site position. If we let the PWM M have entries $M_{c,j} = \log(f_{c,j}/p_c)$, it is easy to see that

$$\log\left(\frac{\Pr(s_i|x_i)}{\Pr(\overline{s_i}|x_i)}\right) \approx \sum_{j=1}^{w} M_{b,j} = S(x_i), \quad (10)$$

where $b = x_{i,j}$ refers to the DNA base in the j-th position of site x_i. In our implementation of the log-posterior odds score, we actually score both DNA strands independently, and our notation could be easily extended (with some loss of clarity) to include this complication (e.g. x_i^+ and x_i^- to indicate the corresponding sites on the two DNA strands).

We estimate the log-prior odds [second term in Equation (9)] using $f(y_i)$, our estimate of $\Pr(B_i|y_i)$, the prior probability of any TF binding to position i. Recall that the Boolean variable B_i is true if genomic region x_i is bound by *some* TF in a given tissue or cell type or condition. We note that if B_i is false, then s_i must also be false, since B_i being false means that *no* TF binds region x_i, that is, $\Pr(s_i|\overline{B_i}, y_i) = 0$. To estimate $\Pr(s_i|y_i)$, we make the further simplifying assumption that the epigenetic data y_i adds no knowledge if we already know the value of the general binding propensity variable B_i. That is, we assume that $\Pr(s_i|B_i, y_i) = \Pr(s_i|B_i)$. We further assume that the probability of a position being the site of a *given* TF, conditional on it being the site of *some* TF, is a constant. That is, we assume that $Pr(s_i|B_i) = \alpha$ for some $0 \leq \alpha \leq 1$. Using the above, we can write the required TF-specific as a scaled version of the non-TF-specific prior,

$$\begin{aligned}\Pr(s_i|y_i) &= \Pr(s_i|B_i, y_i)\Pr(B_i|y_i) + \Pr(s_i|\overline{B_i}, y_i)\Pr(\overline{B_i}|y_i) \\ &= \Pr(s_i|B_i, y_i)\Pr(B_i|y_i) \\ &= \Pr(s_i|B_i)\Pr(B_i|y_i) \\ &= \alpha\Pr(B_i|y_i).\end{aligned}$$

Typically, the epigenetic data y_i will not be specific to a given TF, but will rather describe the general binding propensity of genomic region x_i. In practice, we find that the value of the scale factor α has little effect on the accuracy of TFBS predictions made by the log posterior-odds score, so we set $\alpha = 1$ in the remainder of this work. We also find that the accuracy is little affected by the value of β over a wide range of values (Section 6 in Supplementary Material), and we use $\beta = 10000$ unless otherwise noted. Note also that if our smoothed auxiliary data y_i is specific to the given TF, that is, $Pr(s_i|y_i) = Pr(B_i|y_i)$, then we can simply set $\alpha = 1$. In any case, with $\alpha = 1$, and using $f(y_i)$ as our estimate of $\Pr(B_i|y_i)$, we now see that the second term in Equation (9) is

$$\begin{aligned}\log\left(\frac{\Pr(s_i|y_i)}{\Pr(\overline{s_i}|y_i)}\right) &= \log\left(\frac{\alpha\Pr(B_i|y_i)}{1-\alpha\Pr(B_i|y_i)}\right) \\ &\approx \log\left(\frac{f(y_i)}{1-f(y_i)}\right) \\ &= P(i). \quad (11)\end{aligned}$$

Since $\hat{S}(x_i) = S(x_i) + P(i)$, combining Equations (10) with (11) shows that the log-posterior odds score is approximately equal to the log-posterior odds, as required.

2.3 Epigenetic datasets for creating priors

To create tissue-specific priors for predicting TF binding, we use a variety of epigenetic datasets. Specifically, we consider DNase I sensitivity, which correlates with transcriptional competence, as well as the histone marks H3K4me1, H3K4me3, H3K9ac and H3K27ac, which have been shown to correlate with active transcription (Kurdistani and Grunstein, 2003). In addition, we use the histone mark H3K27me3 as a negative control, given that is a key marker of epigenetic transcriptional repression. The sources of the data, as well as the tissue types and histone modifications we study, are summarized in Supplementary Table S2.

In our first experiment, to directly compare log-posterior odds motif scoring with our previous work (Whitington *et al.*, 2009), we use H3K4me3 ChIP-seq data from mouse embryonic stem (ES) cells (Mikkelsen *et al.*, 2007). These data represent the 'density' of antibody-enriched fragments, calculated at 25 bp resolution. Each value is calculated by adding one to the count for each uniquely aligned fragment occurring within 200 bp of the given position. Reads occurring 200–300 bp from a given position contribute a count of 0.25 to the count in that 25 bp interval. We use this count as the value of y_i in Equation (1) for all positions in the 25 bp window, and we replace β with $\beta/25$ in that equation and in Equation (2) since there are now only $n/25$ prior values that must sum to β.

In our other experiments, we use ChIP-seq histone mark datasets from two human cell lines—K562 and GM12878. These data were generated by the Bernstein lab as part of the ENCODE Consortium (Myers *et al.*, 2011), where the raw tag counts were converted to densities as described in the preceding paragraph. In these experiments, for the K562 cell line we use ENCODE DNase I hypersensitivity (HS) data generated by the Stamatoyannopoulos lab at the University of Washington. These data are provided as smoothed data that specifies the total number of tags mapping to a genomic window of 150 bp, with windows specified every 20 bp (Myers *et al.*, 2011). In order to compare our results with previous work by others (Pique-Regi *et al.*, 2011), for the GM12878 cell line we use ENCODE DNase I HS data from the Crawford lab at Duke University (Boyle *et al.*, 2011). We converted these data to exactly the same format as the K562 data, smoothing tag counts in a window of 150 bp with windows every 20 bp. We use this density as y_i in Equation (1) for all positions in the 20 bp window, and we replace β with $\beta/20$ in that equation.

In order to study a novel prediction approach based on filtering by PWM score followed by sorting on DNase I HS, we also prepared an unsmoothed version of the same GM12878 DNase I HS data. We processed the tagAlign files from the three replicate files separately, discarding multi-mapping tags, and collapsing all duplicate counts within a single file to a count of 1. We then pooled the processed data from the replicates, creating a file containing the total number of tags at each genomic position.

2.4 Defining TFBS gold standards based on ChIP-seq data

We used two types of gold standard for evaluating TFBS prediction methods. For a given TF and tissue, each gold standard is created by considering ChIP-seq data for that TF in that tissue as well as a known PWM for the TF. These PWMs are described in Supplementary Table S3.

Our first two experiments are carried out using a 'peak-centric' gold standard, which defines a single position in each ChIP-seq peak as a binding site for the TF (a positive), and treats all other genomic positions as negatives. The single positive position assigned to each ChIP-seq peak is the position with the highest PWM score. No positive site is assigned to peaks that do not contain a site with a $P < 0.005$, as computed by FIMO (Grant *et al.*, 2011). Using this gold standard allows us to compare with the earlier work of Whitington *et al.* (2009). The ChIP-seq datasets we use for generating the peak-centric gold standards are summarized in Supplementary Table S1.

In our third experiment, to compare our approach against a recently published method, we use the 'site-centric' gold standard that was used by Pique-Regi *et al.* (2011). This gold standard first removes from consideration all genomic positions that are not *potential* sites, and labels as 'positive' and 'negative' the sites with strong ChIP-seq evidence for or against occupancy by the TF in the given tissue. Potential sites are those with PWM scores greater than $\log_2(10000) = 13.28$ bits. Positive sites are potential sites that occur within a declared ChIP-seq peak. Negative sites are sites that are not within a ChIP-seq peak and the ChIP-seq control signal (tag count) is higher than the ChIP-seq signal. All potential sites that are not clearly positives or negatives according to these two rules are removed from consideration.

2.5 TFBS prediction accuracy metrics

We use two metrics for measuring the accuracy of predictions. Both are based on the true positive rate (TPR) and the false positive rate of a predictor. The TPR, which is often called sensitivity, is the number of true positives divided by the number of known positives at a given score threshold. The FPR is the number of false positives divided by the number of known negatives at a given score.

The first accuracy metric we use is the area under the receiver operating characteristic curve (AUC). This metric gives a combined measure of accuracy at all sensitivity rates, by integrating TPR over all values of FPR (Swets, 1988). The second metric we use is sensitivity at 1% FPR, which allows us to compare the accuracy of prediction methods when a premium is placed on avoiding false positives.

3 RESULTS

3.1 H3K4me3 prior improves TFBS prediction accuracy in mouse ES cells

In our first experiment, we use the log-posterior odds score with the H3K4me3 prior to predict the genome-wide binding of 13 TFs in mouse ES (mES) cells. This experiment tests the hypothesis that our probabilistic prior yields improved TFBS prediction accuracy relative to two competing methods: using no prior at all, or using a 'hard' prior, in the form of a site filter based on histone modification ChIP-seq tag counts. We described the latter approach previously (Whitington *et al.*, 2009); accordingly, this experiment uses the same experimental design as that previous work. Specifically, we employ peak-centric gold standards, as described in Section 2, prepared using the TF ChIP-seq datasets listed in Supplementary Table S1. The 13 TFs are cMyc, CTCF, E2f4, Esrrb, Klf4, Nanog, nMyc, Oct4, Smad1, Sox2, Stat3, Tcfcp2l1 and Zfx.

The results shown in Figure 1 demonstrate that predictions made using the log-posterior odds score are substantially more accurate than those made using the PWM score. The sensitivity at 1% FPR

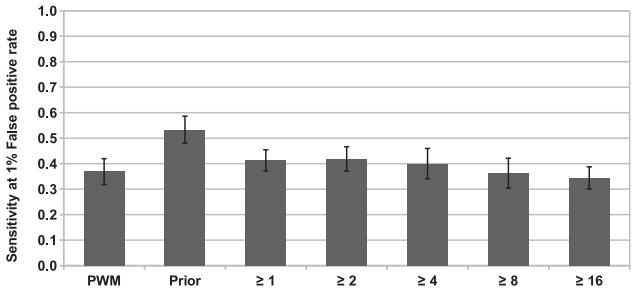

Fig. 1. Accuracy of H3K4me3 log-posterior odds score compared with H3K4me3 filtering in mES cells. Results are shown for predicting the binding sites of 13 TFs in mES cells. Accuracies are measured using peak-centric gold standards. The solid bars represent the mean sensitivity at 1% false positive rate; error bars show standard error. 'PWM' refers to using just the PWM score; 'Prior' refers to the log-posterior odds score using the H3K4me3 prior; '$\geq n$' refers to the H3K4me3 histone-filtering method using a tag-count threshold of n.

improves from an average of 37% using the PWM score to over 53% using the H3K4me3 prior. As indicated by the small standard error bars, accuracy using the prior is fairly consistent for all 13 TFs, ranging from 77% for Nanog to 98% for CTCF (data not shown). In the Supplementary Material, we show that the log-posterior odds score is also substantially more accurate than the PWM score when we measure the FPR at a given sensitivity (Supplementary Fig. S3 and Table S6). For example, at 20% sensitivity, using the log-posterior odds score reduces the FPR by ~68%.

In this experiment, the H3K4me3 log-posterior odds score is also substantially more accurate than our previous method of filtering by histone score and sorting by PWM score (Whitington *et al.*, 2009). The best filtering approach ($n \geq 2$) has a average sensitivity of only 42%, compared with 53% using the log-posterior odds score (Fig. 1). The H3K4me3 log-posterior odds score is also more accurate than the filtering approach in terms of FPR at sensitivity levels between 20% and 100% (Supplementary Fig. S3).

The H3K4me3 log-posterior odds score also achieves higher overall AUC than either the PWM score or any of the filtering methods tested. (Supplementary Fig. S2). The average AUC for predicting the binding sites of the 13 TFs is 92.4% for using the log-posterior odds score. The best filtering approach we test ($n \geq 2$) has average AUC of 91.5%, and the PWM score achieves 88.6% AUC. The small differences in AUC among the different methods shows that AUC obscures the significant differences at low false positive rates. This is due to the large imbalance in the number of positives and negatives in the peak-centric gold standard causing AUC to emphasize accuracy at high false positive rates.

3.2 Histone modification and DNase I HS priors improve TFBS prediction in human K562 cells

In our second experiment, having established the utility of our log-posterior odds scoring in mouse, we perform a systematic evaluation of a variety of priors in human K562 cells. Our aims in this experiment are 2-fold: first, to verify that our proposed method works well with priors derived from a variety of types of histone modifications or from DNase I sensitivity data; second, to examine the extent to which multiple priors can be combined to achieve even better motif search performance. For evaluation, we again

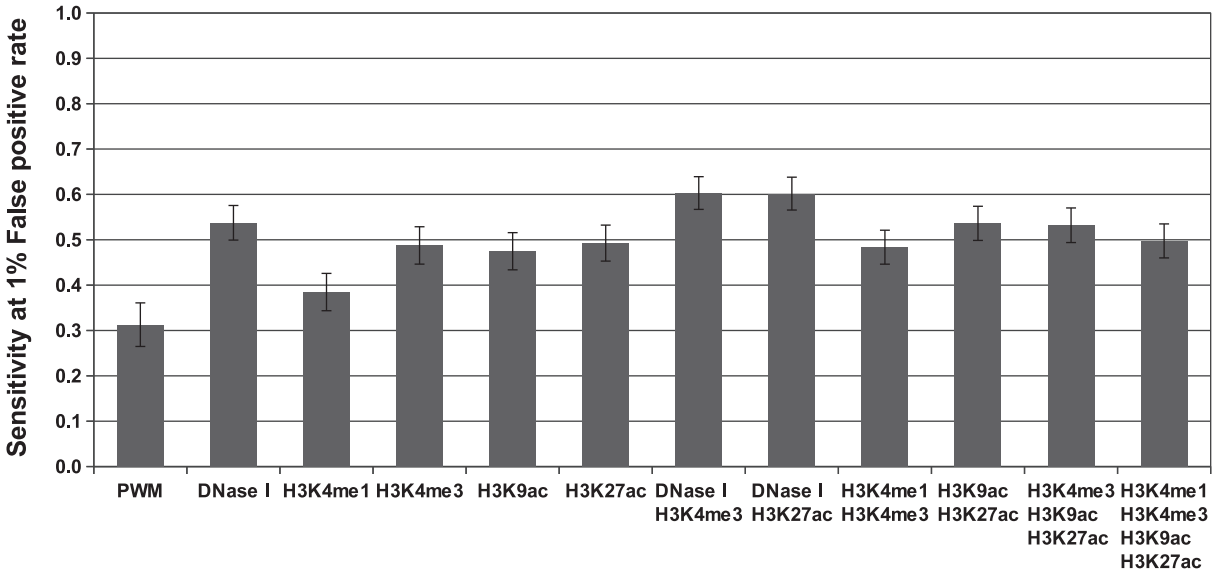

Fig. 2. The log-posterior odds score based on various histone marks and DNase I data improves binding site recognition in human K562 cells. Results are shown for predicting the binding sites of 15 TFs in K562 (human erythroleukaemia) cells. The height of each bar corresponds to the average sensitivity at 1% false positive rate, and error bars indicate standard error. All DNase I hypersensitivity data are from the Stamatoyannopoulos lab at the University of Washington.

use a peak-centric gold standard, this time derived from 15 TFs (Supplementary Table S1). The priors used in this second experiment are based on the histone ChIP-seq and DNase I datasets for K562 cells listed in Supplementary Table S2. The 15 TFs are Atf3, Brg1, cFos, cMyc, Egr1, Gabp, Gata1, Jund, Max, Nfe2, Nfya, Nfyb, Sirt6, Usf1 and YY1.

The results of this experiment, summarized in Figure 2, show clearly that log-posterior odds scoring is much more accurate than the PWM score. For example, for 15 TFs the average sensitivity at 1% FPR is 53.7% using the DNase I prior, compared with only 31.3% using PWM scanning. Among the five priors based on single epigenetic datasets used in Figure 2 (DNase I HS, H3K4me1, H3K4me3, H3K9ac and H3K9ac), the DNase I prior is the most informative according to this accuracy metric and according to AUC (Supplementary Fig. S4). Strikingly, at both 20 and 50% sensitivity, the log-posterior odds score with the DNase I HS prior reduces the number of false positives by almost 90% on average when predicting the binding of these 15 TFs (Supplementary Fig. S5). Among the four single-histone priors, H3K27ac and H3K4me3 are the most informative, achieving sensitivities of 49.3 and 48.8%, respectively, at 1% FPR. This result agrees with our previous observation that H3K4me3 is highly predictive of TF binding compared with other histone marks (Whitington *et al.*, 2009). The prior based on H3K4me1 data is much less effective in this experiment, yet still achieves a higher sensitivity rate (38.5%) than PWM scanning (31.3%). These results suggest that DNase I sensitivity data is most useful in predicting TF binding, but that in the absence of DNase I HS data, priors based on H3K27ac, H3K4me3 or even H3K4me1 ChIP-seq data can still greatly increase the accuracy of TFBS prediction when used with the log-posterior odds score.

Finally, we investigate if multiple priors can be combined to achieve even greater prediction accuracy. We observe in Figure 2 that combining DNase I sensitivity data with an informative histone modification (H3K4me3 or H3K27ac) yields an improvement in sensitivity, relative to using any of the three priors alone. For example, the sensitivity of the log-posterior score using a composite prior based on DNase I and H3K4me3 data is 60.3% at 1% FPR, compared with 53.7% using the DNase I prior or 48.8% using the H3K4me3 prior alone.

Somewhat surprisingly, we found that using multiple histone marks to construct priors did not improve accuracy greatly compared to using a single histone prior. For example, a prior constructed from all four of our (non-control) histone marks—H3K4me1, H3K4me3, H3K9ac and H3K27ac—achieves sensitivity of 49.7% at 1% FPR, only marginally better than the accuracy of the prior based on H3K27ac alone (49.3% sensitivity at 1% FPR). However, a prior constructed from H3K9ac and H3K27ac does show moderately improved accuracy (Fig. 2). When accuracy is measured in terms of AUC, the conclusions are the same (Supplementary Fig. S4).

The failure of the combined histone priors to improve over the individual priors may be due to the fact that these three histone marks are highly correlated (Supplementary Fig. S7). On the assumption that the lack of improvement was due to correlated priors not contributing additional information, we also tested a prior combining H3K4me1, which is far less correlated with the other three, with the best histone prior, H3K4me3. However, this prior—H3K4me1 and H3K4me3—performs no better than the K3K4me3 prior alone (Fig. 2).

3.3 DNase I-based scores dramatically improve TFBS prediction in human GM12878 cells

In our third experiment, we compare the log-posterior odds score method to a recently described method called CENTIPEDE (Pique-Regi *et al.*, 2011). Accordingly, this experiment follows the design used by Pique-Regi *et al.* (2011). Specifically, we use site-centric

Table 1. Performance of five TFBS prediction methods on putative sites with high PWM score

TF	Neg	Pos	Area under the ROC curve (%)					Sensitivity at 1% false positive rate				
			PWM	H-p	D-p	Cpd	D-s	PWM	H-p	D-p	Cpd	D-s
CTCF	23201	17093	80.2	82.5	93.6	97.5	**97.9**	26.4	27.4	44.5	**87.6**	33.7
Gabpa	19150	924	61.4	96.4	99.1	99.7	**99.8**	2.0	27.5	70.2	95.0	**98.1**
Jund	72172	14	82.0	92.7	**99.4**	98.6	99.0	21.4	0.0	**78.6**	**78.6**	57.1
Max	13611	72	70.6	93.3	94.6	99.6	**99.7**	8.0	36.1	33.1	88.9	**94.4**
Nrsf	1642	1193	89.1	88.1	92.5	**95.2**	89.3	57.1	55.3	**62.7**	46.2	13.9
Srf	11146	133	72.3	90.6	95.4	99.2	**99.7**	12.0	25.5	42.8	92.5	**93.2**
Mean			75.9	90.6	95.8	**98.3**	97.6	21.2	28.6	55.3	**81.5**	65.1

The table shows two measurements of accuracy of tissue-specific TFBS prediction in GM12878 (lymphoblastoid) cells. The first three columns list the transcription factor (TF) and numbers of negative (Neg) and positive (Pos) sites in the site-centric gold standard for that TF. Accuracies (area under the ROC curve and sensitivity at 1% false positive rate) are shown for sorting sites by either the PWM score (PWM), the log-posterior odds using the H3K4me3 prior (H-p), the log-posterior odds using the DNase I prior (D-p), the CENTIPEDE log-posterior odds (Cpd) or the DNase I score (D-s). All DNase I data are from the Crawford lab at Duke University. Best results for each accuracy metric are shown in bold.

gold standards for six TFs in the GM12878 lymphoblastoid cell line. The first three columns of Table 1 list these TFs along with the corresponding numbers of positive and negative sites in the site-centric gold standards. As before, we use as performance metrics the AUC and sensitivity achieved at 1% false positive rate, as used by Pique-Regi *et al.* (2011). We also explored accuracy in terms of sensitivity at 1% false *discovery* rate (Supplementary Table S8), with results essentially identical to those described below.

The results of this experiment, summarized in the left half of Table 1, show two very clear trends. First, as in our previous two experiments, using the log-posterior odds score with an epigenetic prior results in higher TFBS prediction accuracy, compared with traditional PWM scoring, with the DNase I prior outperforming the H3K4me3 histone prior. For example, the mean AUC when predicting the binding of these six TFs in GM12878 cells is 75.9% when putative sites are sorted by the PWM score, 90.6% using the H3K4me3 prior and 95.8% when we use the DNase I prior. Using either of these two priors also results in a large increase in the sensitivity achieved at a 1% FPR. On average, the sensitivity more than doubles when we use the DNase I HS prior, compared with using the PWM score (55.3% versus 21.2% sensitivity). Using the H3K4me3 prior, the improvement in sensitivity at 1% FPR is far less (sensitivity = 28.6%). We also tested other histone priors and combinations of histone priors, and the results are very similar (data not shown).

Second, the CENTIPEDE method, which leverages DNase I HS, sequence conservation and distance from TSS data, usually performs better than the log-posterior odds score using only the H3K4me3 or DNase I prior (Table 1). For all six TFs, CENTIPEDE's AUC is higher than the corresponding AUC for our method using the DNase I prior. Overall, the average difference in AUC is 2.5%, which is relatively large considering that the AUC scores themselves are quite close to 100%. The results with respect to sensitivity at 1% FPR are similar, with CENTIPEDE performing better than the DNase I prior in four out of six cases.

However, while investigating the cause for CENTIPEDE's strong performance in this experiment, we made a surprising discovery: we are able to achieve performance comparable to CENTIPEDE's by completely ignoring the PWM and using only the DNase I data. This strawman method, which we call the 'DNase I score', simply sorts all putative sites by the total number of DNase I cuts (tags) in a window of 150 bp around the site. Recall that the site-centric gold standards contain only putative sites with PWM scores greater than $\log_2(10000) = 13.28$ bits. Hence, the DNase I score is analogous to the 'histone filter' proposed by Whitington *et al.* (2009), except that rather than filtering by histone tag counts and ranking by PWM score, the DNase I score filters by PWM score and ranks by DNase I tag counts. As can be seen from Table 1 (column 'D-s'), the DNase I score AUC is higher than CENTIPEDE's (and than the DNase I prior's) for five out of six TFs. For the remaining TF (Nrsf), CENTIPEDE achieves a much higher AUC (95.2% compared with 89.3%). In terms of sensitivity at 1% FPR, the DNase I score performs substantially better than CENTIPEDE for three of the six TFs, but is much worse for the other three TFs: Nrsf, CTCF and Jund. These results suggest that the DNase I score method is as accurate as CENTIPEDE in terms of AUC, but less accurate at very low false positive rates.

The above observation—that simply ranking sites based on DNase I sensitivity yields results that are competitive with the best-performing method—suggests that, if we only consider sequences whose PWM scores are relatively high, DNase I sensitivity is more useful than the DNA sequence at discriminating between bound and unbound sequences. It also makes clear a weakness in the site-centric gold standard. Since the site-centric gold standard is based solely on sites with PWM scores exceeding a given threshold, it cannot tell us anything about the accuracy of a prediction method on sites with PWM scores below the threshold. Other types of gold standards, such as the 'peak-centric' one used here, must be employed in order to compare the ability of prediction methods on more difficult binding sites with lower PWM scores.

4 DISCUSSION

We have demonstrated that the log-posterior odds score is a powerful method for improving tissue-specific TFBS prediction when genome-wide histone modification ChIP-seq or DNase I hypersensitivity data is available. Compared to a sequence-only scan or to using a hard filter based on histone or DNase I HS tag counts,

the log-posterior odds score offers a large and consistent advantage. Among the datasets that we considered, DNase I sensitivity provides the most useful prior, though some additional benefit can be gained by combining priors from DNase I and an informative histone modification such as H3K4me3 or H3K27ac.

When we compare our proposed method to CENTIPEDE, which leverages additional sources of information such as distance from transcription start sites, the conclusions are less clear. On the one hand, when measured via AUC or sensitivity at 1% FPR, CENTIPEDE appears to provide slightly better results than the log-posterior odds score. On the other, we have shown that, for this particular benchmark, simply ranking sites by their smoothed DNase I tag counts performs comparably to CENTIPEDE. This observation suggests that, among sites with high PWM scores, measurements of DNase I sensitivity are much more useful than the PWM score itself in discriminating between bound and unbound sequences. The experiment also leaves open the possibility that, in a more realistic setting (e.g. using a gold standard that includes all sites in the genome), the relative performance of the methods would be quite different. Finally, it is worth noting that the relative simplicity of the proposed log-posterior odds score makes it an appealing alternative to a more complex approach such as CENTIPEDE.

Although our work has focused on scanning for occurrences of DNA motifs, the log-posterior odds score framework can be trivially extended to protein motifs, given a suitable prior. Such priors might be derived, for example, from the measurements of evolutionary conservation or from the predicted local attributes such as signal peptides, solvent accessibility, transmembrane topology or propensity to bind to DNA or to other proteins. Similarly, the method described here could be easily incorporated into existing probabilistic approaches for modeling regulatory sequences as *cis*-regulatory modules composed of multiple motifs (Bailey and Noble, 2003; Sinha *et al.*, 2008; Zhou and Wong, 2004). Although our linear mapping function for constructing epigenetic priors, $g(y_i)$, works well in practice, the use of a non-linear function is a promising direction for future research.

ACKNOWLEDGEMENTS

We thank Roger Pique-Regi for providing the site-centric gold standard for GM12878 cells.

Funding: National Institutes of Health award (R01 RR021692).

Conflict of Interest: none declared.

REFERENCES

Bailey,T. L. and Noble,W. S. (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **19** (Suppl. 2), ii16–ii25.

Barski,A. *et al.* (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Bernat,J.A. *et al.* (2006) Distant conserved sequences flanking endothelial-specific promoters contain tissue-specific DNase-hypersensitive sites and over-represented motifs. *Hum. Mol. Genet.*, **15**, 2098–2105.

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

Crawford,G.E. *et al.* (2006) DNase-chip: a high-resolution method to identify DNase I hypersensitive sites using tiled microarrays. *Nat. Methods*, **3**, 503–509.

Cui,K. *et al.* (2009) Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell Stem Cell*, **4**, 80–93.

Duda,R.O. *et al.* (2001) *Pattern Classification*. John Wiley & Sons, New York.

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

Gordân,R. *et al.* (2010) Finding regulatory DNA motifs using alignment-free evolutionary conservation information. *Nucleic Acids Res.*, **38**, e90.

Grant,C.E. *et al.* (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

Heintzman,N. *et al.* (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.*, **39**, 311–318.

Heintzmann,N.D. *et al.* (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, **459**, 108–112.

Hesselberth,J. *et al.* (2009) Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. *Nat. Methods*, **6**, 283–289.

Keene,M.A. *et al.* (1981) DNase I hypersensitive sites in Drosophila chromatin occur at the 5' ends of regions of transcription. *Proc. Natl Acad. Sci. USA*, **78**, 143–146.

Kurdistani,S.K. and Grunstein,M. (2003) Histone acetylation and deacetylation in yeast. *Nat. Rev. Mol. Cell Biol.*, **4**, 276–284.

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

McArthur,M. *et al.* (2001) Quantification of DNaseI-sensitivity by real-time PCR: quantitative analysis of DNaseI-hypersensitivity of the mouse β-globin LCR. *J. Mol. Biol.*, **313**, 27–34.

Mikkelsen,T.S. *et al.* (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.

Myers,R.M. *et al.*; ENCODE Consortium (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*., **9**, e1001046.

Narlikar,L. *et al.* (2006) Informative priors based on transcription factor structural class improve de novo motif discovery. *Bioinformatics*, **22**, e384–e392.

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

Robertson,G. *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods*, **4**, 651–657.

Sinha,S. *et al.* (2008) Systematic functional characterization of cis-regulatory motifs in human core promoters. *Genome Res.*, **18**, 477–488.

Swets,J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

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

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

Wu,C. (1980) The 5' ends of *Drosophila* heat shock genes in chromatin are hypersensitive to DNase I. *Nature*, **286**, 854–860.

Zhou,Q. and Wong,W. (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. *Proc. Natl Acad. Sci. USA*, **101**, 12114–12119.