Bioinformatics, 31(17), 2015, 2852–2859
doi: 10.1093/bioinformatics/btv294
Advance Access Publication Date: 7 May 2015
Original Paper

Systems biology

BinDNase: a discriminatory approach for transcription factor binding prediction using DNase I hypersensitivity data

Juhani Kähärä* and Harri Lähdesmäki

Department of Information and Computer Science, Aalto University School of Science, FI-00076 Aalto, Finland

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 10, 2014; revised on April 15, 2015; accepted on May 4, 2015

Abstract

Motivation: Transcription factors (TFs) are a class of DNA-binding proteins that have a central role in regulating gene expression. To reveal mechanisms of transcriptional regulation, a number of computational tools have been proposed for predicting TF-DNA interaction sites. Recent studies have shown that genome-wide sequencing data on open chromatin sites from a DNase I hypersensitivity experiments (DNase-seq) has a great potential to map putative binding sites of all transcription factors in a single experiment. Thus, computational methods for analysing DNase-seq to accurately map TF-DNA interaction sites are highly needed.

Results: Here, we introduce a novel discriminative algorithm, BinDNase, for predicting TF-DNA interaction sites using DNase-seq data. BinDNase implements an efficient method for selecting and extracting informative features from DNase I signal for each TF, either at single nucleotide resolution or for larger regions. The method is applied to 57 transcription factors in cell line K562 and 31 transcription factors in cell line HepG2 using data from the ENCODE project. First, we show that BinDNase compares favourably to other supervised and unsupervised methods developed for TF-DNA interaction prediction using DNase-seq data. We demonstrate the importance to model each TF with a separate prediction model, reflecting TF-specific DNA accessibility around the TF-DNA interaction site. We also show that a highly standardised DNase-seq data (pre)processing is a requisite for accurate TF binding predictions and that sequencing depth has on average only a moderate effect on prediction accuracy. Finally, BinDNase's binding predictions generalise to other cell types, thus making BinDNase a versatile tool for accurate TF binding prediction.

Availability and implementation: R implementation of the algorithm is available in: http://research.ics.aalto.fi/csb/software/bindnase/.

Contact: juhani.kahara@aalto.fi

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

1 Introduction

Transcriptional regulation is largely controlled by transcription factors (TFs) that bind short DNA sequence motifs in gene promoters, enhancers and other regulatory sites. Many TFs bind DNA in a sequence specific manner and understanding TF binding is integral to understanding gene regulatory networks. Moreover, changes in the genomic DNA, such as SNPs, at TF-DNA interaction sites can affect TF binding and can contribute to phenotypic differences, including gene expression (Kasowski *et al.*, 2010), but can also contribute to various diseases (Matsuda *et al.*, 1992; Wittwer *et al.*, 2006). Determining the locations of TF-binding sites is therefore of high importance.

Commonly used computational strategies for predicting TF-binding include DNA motif based prediction of TF binding sites

(Weirauch *et al.*, 2013) and simultaneous analysis of DNA motifs and histone modification data (e.g. (Ramsey *et al.*, 2010)) or DNase I hypersensitivity experiments followed by sequencing (DNase-seq) (Boyle *et al.*, 2008). The current state-of-art method for genome-wide profiling of TF-binding is chromatin immunoprecipitation followed by sequencing (ChIP-seq). ChIP-seq can however map the positions of only one TF per experiment and requires a specific, ChIP-grade antibody for the protein of interest. On the contrary, DNase-seq detects signal at open chromatin sites genome-wide. Consequently, DNase-seq is increasingly used to complement ChIP-seq experiments because a single DNase-seq experiment can provide valuable information about putative TF-DNA interaction sites for all TFs. Genome-wide maps of putative regulatory sites in selected cell types detected using DNase-seq data have already been created e.g. in the ENCODE project (Neph *et al.*, 2012).

Currently the exact locations of TF-binding events are pinpointed by finding stereotypic DNase I footprints. These footprints are short genomic locations of low DNase I cleavage activity immediately flanked by high DNase activity. An illustration of such regulatory site is shown in Figure 1a where the ATF1 motif locations within ChIP-seq peaks are characterised with low DNase activity and the flanking regions exhibit high DNase activity. Finding stereotypic DNase I footprints does not require any information about the DNase I signal at the true binding sites, and therefore many methods using this definition of a TF footprint are unsupervised methods. It should also be acknowledged that these algorithms are transcription factor agnostic as they can only predict whether any TF is bound to a given site.

It has however been reported in (Neph *et al.*, 2012), that for some proteins nucleotides in the middle of TF-DNA interface are exposed to DNase I cleavage. Therefore, treating all the nucleotides as protected in the TF-DNA interface might not be adequate. The exposed nucleotides differ between TFs and this high resolution information might be useful for making more accurate predictions and distinguishing which TFs actually occupy the sites. Most of the currently used methods developed for identifying footprints use the canonical definition of DNase footprints of low DNase activity flanked by high activity. These methods can be supervised (MILLIPEDE) (Luo and Hartemink, 2013), in which the predictions are made using models trained on training data, or unsupervised (Neph *et al.*, 2012; Piper *et al.*, 2013), which rely on giving TFs binding scores according to a model devised without training data. Some unsupervised methods, such as CENTIPEDE and PIQ, include nucleotide resolution information (Pique-Regi *et al.*, 2011; Sherwood *et al.*, 2014).

On the other hand, a recent paper shows that the nucleotide resolution DNase I cleavage pattern is partly caused by the intrinsic sequence bias of the DNase molecule (He *et al.*, 2014) suggesting that the nucleotide resolution DNase-seq signal at the TF-DNA interaction site does not necessarily provide predictive power to distinguish real binding sites. Moreover, DNase I footprint signal at individual genomic locations are noisy as illustrated in Figure 1b. It is therefore difficult to tell which signals in the DNase-seq data are informative about TF binding and which signals are either noise or caused by the intrinsic sequence bias. Consequently, carefully designed computational methods are needed for DNase-seq data processing.

Having the aforementioned characteristics of DNase-seq data in mind, here we study the use of high resolution DNase-seq data for predicting TF binding sites. We chose to treat the TF binding prediction problem as a supervised classification task, because this approach lets us utilize the true differences in the DNase signal between the bound and unbound sites in the model training. We develop a method, BinDNase, which for each TF automatically extracts features from the DNase-seq data which maximally discriminate bound and unbound genomic locations. By comparing BinDNase with state-of-the-art TF binding prediction methods we show that BinDNase provides more accurate predictions than other methods. Although this study focuses on making predictions with DNase-seq data, the method is readily applicable to other data types such as FAIRE-seq (Giresi *et al.*, 2006) or ATAC-seq (Buenrostro *et al.*, 2013).

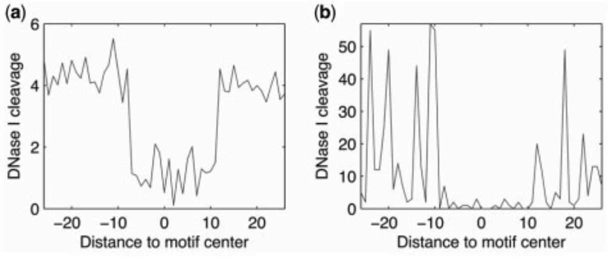

Fig. 1. (a) The average DNase I cleavage around ATF1 binding sites resembles the canonical definition of DNase I footprint. The average DNase-seq signal at nucleotide resolution centered at ATF1 motif overlapping ATF1 ChIP-seq peaks is shown. **(b)** DNase-seq signal at nucleotide resolution around a single AFT1 binding site located between nucleotides 96 929 096–96 929 148 in chromosome 9. Although individual footprints are noisy, the canonical shape of the footprint is still visible in the data

2 Materials and methods

2.1 The data

This work utilizes the publicly available data from the ENCODE project (The ENCODE Project Consortium, 2012). The digital genomic footprinting datasets (track name UwDgf) were obtained from the ENCODE website. The datasets include raw reads, DNase I signal and DNase I hotspots. The DNase I signal is a genome wide map that lists the number of reads starting from each nucleotide position. Note that the 5′ end of each aligned read and the previous nucleotide to 5′ direction define the DNase I cleavage sites. The hotspots are genomic regions of high DNase I activity as identified by the hotspot-algorithm (John *et al.*, 2011).

Candidate binding sites were found for each TF by scanning human genome (version hg19) with the position specific frequency matrix (PSFM) models from the ENCODE publication (Wang *et al.*, 2012). The motif matches were detected with p-value of 10^{-5} using the motif scanning software FIMO. The motif matches were divided into three categories:

- Motif within a ChIP-seq peak = Truly bound sequence (positive set)
- Motif not within a ChIP-seq peak = Unbound sequence (negative set 1)
- Motif not within a ChIP-seq peak but within a hotspot region = Unbound sequence (negative set 2)

The ChIP-seq peaks and hotspot regions were downloaded from the ENCODE website. The list of the ENCODE files that were used can be found in the supplemental material. We use a combination of positive and negative set (either 1 or 2) to train our discriminative model, BinDNase. At the beginning all data sets are split into two parts: 200 positive sites and 1000 negative sites are chosen randomly to the test data sets, rest of the data (but at maximum 3000 positive or negative sites) is used in the training data sets. The number of sites in the test set is relatively low because we wanted to have a comparable number of sites for each TF. The number of sites does not however alter the results considerably (Supplemental Information 6). The final prediction accuracy evaluation of the models is conducted using the test data which is not used to train the model.

2.2 Logistic regression

In this work the candidate binding sites found with PSFM modeling are scored according to the DNA cleavage data from DNase-seq experiments. DNase I signal is modelled using the standard logistic regression where the log-transformed counts of DNase I induced cuts are the input variables. We use the DNase-seq data D at multiple resolutions: At single nucleotide resolution the input variable value D_i represents the log-transformed number of DNase I induced cuts at that particular nucleotide. For lower resolutions the values of the input variables correspond to the log-transformed aggregate number of reads falling into bins of varying size $\ell > 1$ (i.e. ℓ consecutive nucleotides). In the following, we use nucleotide and lower resolution data interchangeably and call them as bins.

More specifically, the probability for binding for each candidate binding site is modelled as follows

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \sum_i \beta_i D_i + \beta_{\mathrm{PSFM}} S_{\mathrm{PSFM}}, \qquad (1)$$

where p is the probability of binding, D_i is the log-transformed number of reads (i.e. DNase I induced cuts) in the bin i, S_{PSFM} is the negative log p-value from FIMO motif scanning and βs are the coefficients in the model. The coefficients are found with the iteratively reweighed least squares (IRLS) algorithm (MATLAB function glmfit). Similar logistic regression model for DNase-seq data has been used in (Luo and Hartemink, 2013)—the main difference is the way the features (i.e. the bins) are chosen for the model, which we describe next.

2.3 Selecting discriminatory features from DNase-seq data

In this work a greedy backward search type machine learning method is implemented to find (and extract) optimal features from the DNase I data for each TF. In the beginning of the search the nucleotides within the candidate binding site and the flanking 10 bp regions are treated in one nucleotide resolution, and the more distal nucleotides are initially in ten nucleotide wide bins. These initial bins are then merged in the search to get more predictive power. In each iteration step two adjacent bins are merged. A schematic presentation with two iterations is presented in Figure 2.

In each iteration the algorithm performs the particular bin merging operation that leads to the best prediction performance. The prediction performance is evaluated using cross-validation by dividing the training data (both positive and negative sets) in two sets. The model is trained using one of the sets and the model performance is evaluated using the other set by making predictions and calculating the area under the curve (AUC) metric. This cross-validation step is conducted 30 times for each bin merging operation to reduce variation caused by random division of the training data into training and test sets within the cross validation in the model training.

3 Results

3.1 Standardized DNase-seq data preprocessing is prerequisite for single nucleotide level analysis

As in all modelling the data should be preprocessed in a standardized way. Previous articles have proposed slightly different processing steps for DNase-seq data (see e.g. Piper *et al.*, 2013; Neph *et al.*, 2012). In the DNase I data special attention should be paid on the following processing steps:

- Reverse strand reads should be shifted 1 bp to 5′ direction (or alternatively forward strand reads should be shifted 1 bp to 5′ direction). This shifting acknowledges the fact that DNase I cleaves the DNA between two consecutive nucleotides. With a single base pair shift on either of the strands the DNA cut sites are contributed consistently to a single nucleotide.
- Orientation of TF binding motifs should be taken into account.

Discrepancies in these steps lead to differences in the data and therefore affect any downstream analysis. For example, some of the digital DNase I data sets which are available on the ENCODE project page have preprocessing differences between cell types. Supplementary Figure 49 shows the average DNase I cleavage profiles for protein JUN in four ENCODE cell types: NHDF-Ad, SKMC, K562 and HepG2. We noticed that reverse strand reads are not consistently shifted between these four cell types (compare Supplementary Figs 49a–b with c–d). While the pattern of DNase hotspot remains practically unaffected in K562 and HepG2 the nucleotide resolution cleavage patterns in the JUN interaction site is clearly distorted due to non-standardized data preprocessing. Consequently, if these discrepancies are not properly corrected, TF binding prediction methods which use high resolution DNase-seq data, such as BinDNase, fail to generalize between cell types. Starting from the raw reads and reprocessing the data carefully using a unified preprocessing pipeline makes differences between cell types disappear (Supplementary Fig. 49 e–f). We observe a similar effect for many TFs (another example is Supplementary Fig. 50 (SRF)). We conclude that careful and highly standardized data preprocessing is essential for detailed analysis of DNase-seq data.

3.2 DNA binding should be modelled separately for each TF

We observed that the selected features as well as the actual prediction models differ greatly between TFs. Two of the models are visualised in Figure 3. First, for some TFs, such as ATF1 (Fig. 3) and SP1 (Supplementary Fig. 23), the canonical definition of DNase I footprint is adequate as the feature selection algorithm finds a model in which the reads falling in the central bins decrease the binding score and the reads falling in the flanking nucleotides increase the binding score.

We also observed that not all strong single nucleotide cleavage patterns present in data are important for discriminating true TF binding sites from random unbound motif locations. For example, NRF1 protein has three evident single nucleotide resolution cleavage sites but only the two peaks flanking the actual binding site are associated with a high positive regression weight by the feature selection algorithm (see Fig. 3b). Note that reads falling to the flanking region on right are weighted more than the single-nucleotide resolution patterns. Thus, some of the intricate patterns in DNase-seq data seem unimportant for discriminating between real TF binding and noise and may represent non-idealities and biases in the data, such as the DNase sequence biases. Thus, efficient feature selection is needed to

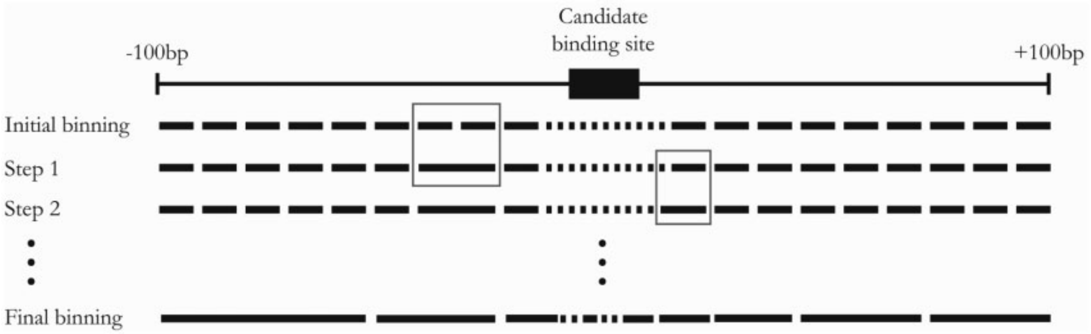

Fig. 2. A schematic presentation of how the optimal features from DNase-seq data are selected. In the initial stage the candidate binding site and 10 bp flanking regions are modeled using 1 bp resolution. The flanking regions 11–100 bp up and downstream from the candidate binding motif are modeled using 10 bp bins. In each step two adjacent bins are merged. In this figure the first step of the algorithm merges two 10 bp bins upstream of the binding site. In the second step one 1 bp wide bin is combined with one 10 bp bin. The final binning in this illustration includes wide bins at the flanking regions and narrower bins at the binding site

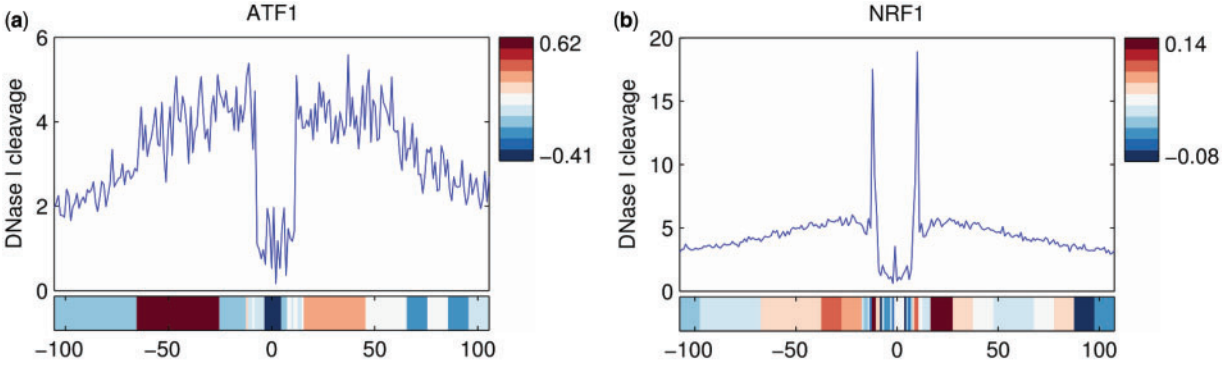

Fig. 3. The upper panels show the average DNase I cleavage centered at the TF binding motifs. The coloured bars indicate the optimised feature selection and their coefficients in the logistic regression model. Red (resp. blue) colour indicates positive (resp. negative) coefficient. **(a)** ATF1, **(b)** NRF1

identify the relevant DNase I cleavage patterns. On the other hand, some of the single nucleotide resolution patterns are highly informative. For example, the DNase I signal for protein CTCF (Fig. 10d) contains a highly stereotypical peak on the right hand side of the binding site. In the optimised feature selection this nucleotide is treated as a single nucleotide and the coefficient in the logistic regression model is high. Moreover, the cleavage pattern in this exact position relative to CTCF motif has previously been reported to differ from the intrinsic sequence bias of the DNase I molecule (He *et al.*, 2014). Models for other TFs, such as ELF1, RAD21, SMC3, SPI1, ETS1 also include a high regression weight coinciding high-resolution DNase pattern (Supplementary Figs 7, 12, 20, 37 and 40). Taken together, these results emphasize that, instead of using a single TF footprint definition, the prediction models for each TF should be constructed separately.

3.3 High resolution DNase-seq analysis improves TF binding predictions

DNase I hypersensitivity at a lower resolution (i.e. for larger genomic regions) has previously been used to detect TF binding sites (see e.g. He *et al.*, 2014; Yardimci *et al.*, 2014). The binding score of each candidate site is then simply the aggregate count of DNase I cuts in that larger window. Reportedly this simple approach performs equally or better than more sophisticated predictors (Yardimci *et al.*, 2014). We evaluated the predictive power of such a lower resolution DNase I activity method using a 50 bp window around the candidate binding site and compared that with BinDNase.

In Figure 4a the TF binding prediction performance is evaluated using candidate sites from random genomic locations (negative set 1) and in Figure 4b using only candidate sites from the hotspot regions (negative set 2). For the less challenging task of discriminating real binding sites from non-binding sites which are not in hotspot regions, we observe that for many TFs it is sufficient to just quantify general DNase I activity near the binding site without using any sophisticated modeling (Fig. 4a).

However, there are TFs for which BinDNase improves the binding prediction accuracy already in this scenario, including e.g. MAFK and ZNF274. The prediction models for MAFK and ZNF274 are shown in Supplementary Figures 16 and 21, which suggest that the relatively poor performance of the simple DNase I activity predictor can be explained by low average DNase I signal at these motifs. BinDNase, in turn, can identify discriminatory features from DNase-seq data and increase the AUC score for MAFK and ZNF274. There are also other TFs that bind to lower DNase activity sites and for which BinDNase improves the prediction results. In general, the amount of improvement in BinDNase's predictions is inversely correlated with the accessibility (i.e. the total cut count) of the region (Supplementary Fig. 1).

Typically TF binding sites are primarily searched for in DNase I hotspot regions and, thus, performance evaluation using the negative set 2 is in practice more relevant. Results in Figure 4b show that while the DNase I activity predictor still works for some TFs surprisingly well it also fails completely for some TFs. The most notable

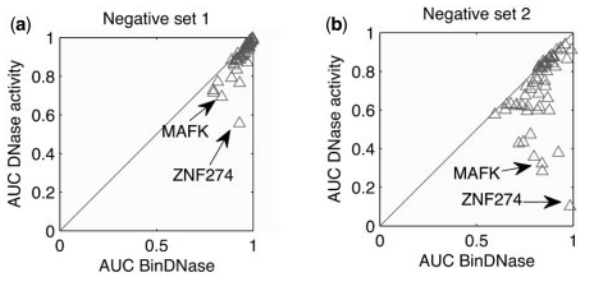

Fig. 4. High resolution DNase-seq analysis improves TF binding predictions when compared with traditional methods. The DNase I activity predictor using a 50 bp window is compared with BinDNase using **(a)** negative set 1 and **(b)** negative set 2. The AUC score is computed for all 57 TFs using both methods. Selected TFs are highlighted in the figure

examples include again MAFK and ZNF274 proteins for which the prediction accuracy is well below the random coin flipping. The worse than random performance can be explained by below-average DNase I activity in the MAFK and ZNF274 binding sites, as the learned BinDNase models consist of mostly wide bins with negative coefficients (Supplementary Figs 16 and 21). As with negative set 1, BinDNase can identify discriminatory features from DNase-seq data and improves AUC scores for majority of the 57 TFs, including MAFK and ZNF274. Interestingly, the prediction models for both MAFK and ZNF274 include also high resolution features, such as a strongly positively weighted single nucleotide features at the middle of the ZNF274 binding site.

3.4 Feature selection improves TF binding predictions

To test if feature selection implemented in BinDNase improves TF binding predictions we also evaluated the prediction performance of logistic regression models which use DNase-seq data only at single nucleotide resolution without implementing any feature selection methods. Results in Figure 5 show that BinDNase's feature selection improves prediction accuracy results for the majority of TFs. The most notable improvement of AUC score is achieved for proteins such as NFYB, TBP and TAF using negative set 2. The models for these TFs consist mostly of wider bins which might indicate that the single-nucleotide model overfits the logistic regression to features that are not useful in discrimination. Without feature selection, some of the nucleotides at the flanking sequences obtain an unnecessarily large weight and hence decrease the prediction accuracy of the nucleotide resolution model. These results suggest that for some nucleotide locations the DNase I signal does not provide single nucleotide resolution information about TF-DNA interaction and those regions should be modelled using larger bins chosen based on a feature selection method. This is expected because the DNase-seq signal at individual candidate binding sites are often noisy and part of the signal originates from the inherent DNase I sequence bias (He *et al.*, 2014). Nevertheless, our results demonstrate that feature selection can identify discriminatory information from DNase-seq data.

We next compared BinDNase with the state-of-the-art methods MILLIPEDE (Luo and Hartemink, 2013) (Fig. 6) and PIQ (Sherwood *et al.*, 2014) (Fig. 7). MILLIPEDE represents the supervised approach for TF binding predictions whereas the PIQ-method is an unsupervised method. The supervised Millipede has been previously shown to outperform the widely used unsupervised CENTIPEDE algorithm (Pique-Regi *et al.*, 2011) for predicting TF binding (Luo and Hartemink, 2013).

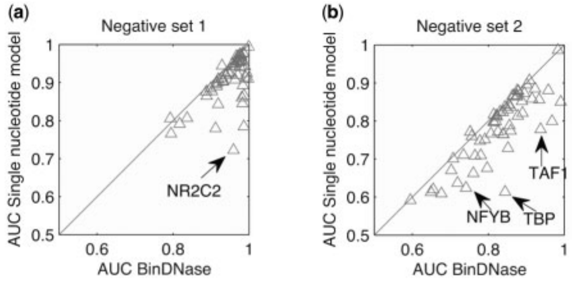

Fig. 5. Feature selection improves TF binding predictions. The AUC scores for BinDNase and a single nucleotide resolution logistic regression model without feature selection for **(a)** negative set 1 and **(b)** negative set 2

BinDNase performs nearly identically with MILLIPEDE for those proteins which are already predicted well by MILLIPEDE. One such protein is ATF1 for which BinDNase finds a prediction model which is close to the canonical footprint model also used in MILLIPEDE (Fig. 3a). However, BinDNase achieves a notable improvement for many of those proteins which MILLIPEDE does not predict well, including e.g. NFYB. Although BinDNase models the flanking regions around the NFYB motif using large size bins similarly with MILLIPEDE, the region close to the NFYB motif center is modeled using very high resolution features with high logistic regression coefficients (Supplementary Fig. 44). BinDNase improves also e.g. CTCF insulator protein whose prediction model is shown in Figure 3d. Previous studies have already indicated a high resolution DNase I cleavage signal on 3′ end of the CTCF motif (He *et al.*, 2014). Here we demonstrate that these high resolution features can be used to improve binding predictions.

Comparison between MILLIPEDE and BinDNase gives expected results because BinDNase can be viewed as a more general version of the MILLIPEDE algorithm: instead of assigning bins similarly for each transcription factor our method finds optimal features for each TF. In the worst case, BinDNase should find similar features and, thereby, similar prediction performance as MILLIPEDE if those indeed happen to be optimal. For other proteins, BinDNase can improve the prediction accuracy as shown in Figure 6b.

Comparison between BinDNase and PIQ methods (Fig. 7) shows that the prediction performance of BinDNase is higher for most TFs. This is also an expected result as the unsupervised PIQ is not designed for the supervised setup, where there is knowledge about the DNase I signal within the true binding sites. This suggests that supervised approaches should be preferred for making accurate binding predictions for TFs for which both ChIP-seq (or other binding information) and DNase-seq data are available for some cell type. Unsupervised methods are required in the absence of accurate ChIP-seq data that is needed in model training or if the ChIP-seq experiment is conducted in a condition where no deeply sequenced DNase-seq data is available.

3.5 Exploring variants of BinDNase algorithm

To further explore the problem of predicting TF binding from DNase-seq data, we devised different versions of the BinDNase for comparison purposes. Motivated by the findings of (Piper *et al.*, 2013), we tried including strand information into BinDNase by counting the DNase induced cuts separately for both strands. The results shown in Supplementary Figure 48 show that the prediction accuracy remains practically the same for negative set 1, whereas the use of the strand information increases the prediction accuracy slightly for selected TFs in the negative set 2. Our analysis shows

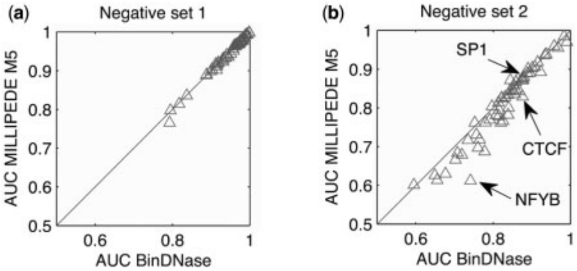

Fig. 6. BinDNase outperforms MILLIPEDE in negative set 2. The AUC scores for BinDNase and MILLIPEDE (Luo and Hartemink, 2013) are presented for **(a)** negative set 1 and **(b)** negative set 2

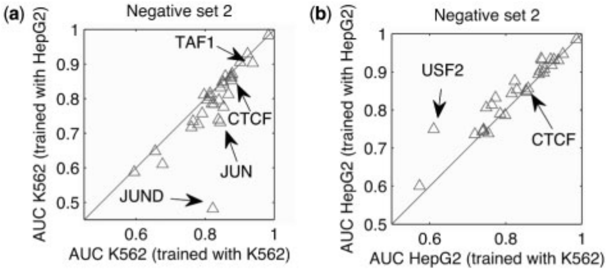

Fig. 8. The models generalize well to different cell types. The model trained on another cell type works almost equally as well as the one trained with the same cell type that the predictions are made to

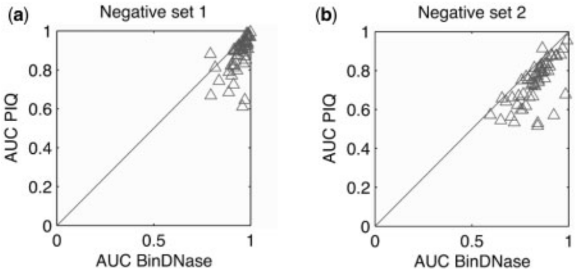

Fig. 7. The prediction performance of supervised BinDNase is higher than the performance of the unsupervised method PIQ for both negative set 1 **(a)** and 2 **(b)**

that strand-specific DNase-seq data may contain some additional information but we consider that the slight improvement does not justify the use of the more complex model.

Following the work in (He *et al.*, 2014; Yardimci *et al.*, 2014), we tested including an explicit sequence bias correction into BinDNase. We chose to model the DNase bias using the 4-mer model from (Sung *et al.*, 2014). The bias signal is further normalized relative to the actual DNase-seq signal in the same region so that the total signal strengths are the same. The estimated bias signal is then subtracted from the original DNase-seq signal and the corrected signal is given as an input to BinDNase algorithm. Surprisingly, our results in Supplementary Figure 47 show that the bias correction performs poorly. A more sophisticated bias correction might work better but that is out of the scope of this work, especially because the current version of BinDNase already outperforms other methods.

We also tested the contribution of the PSFM score in our final predictions by comparing the BinDNase method against a version of BinDNase that does not utilize the motif score. The obtained prediction accuracies in Supplemental Figure 46 show that BinDNase's prediction accuracy decreases only a little in the absence of the motif score. This suggests that DNase signal alone is indeed highly informative of TF binding, although the motif score should be used to achieve the best possible prediction accuracy. Note that the methods used in the comparisons (MILLIPEDE and PIQ) also utilize the motif score.

3.6 BinDNase generalizes between different cell types

To test whether the models generalize between different cell types we used K562 for training the model and made the predictions for cell type HepG2, and vice versa. The models' prediction performances are very similar between cell types for all except a handful of TFs as shown in Figure 8. Naturally, whenever there is a difference in the model performance the predictions made to the same cell type that the model was built with are more accurate.

Some of the proteins that behave similarly and differently between cell types are highlighted in Figure 10. Proteins that do not generalize well belong to e.g. helix-loop-helix (HLH) (BHLHE40) and leucine zipper families (JUND) and USF2 contains both HLH and leucine zipper domains. BinDNase may generalize differently for different structural families. It is interesting to look at the BinDNase models learned from two different cell types in detail. For example, the two models for USF2 (Fig. 10a and c) and JUND (Supplementary Fig. 1a and c) are clearly different as expected due to the weak generalization. The two models for CTCF (Fig. 10b and d) and TAF1 (Supplementary Fig. 28), in turn, have very similar characteristics and, thereby, achieve similar AUC scores. For TAF1, the two models are typical canonical footprint models, whereas for CTCF the models contain both high and low resolution signals, consistent for both cell types.

Overall, despite of small differences in feature extraction (i.e. binning) and regression coefficients for some TFs, it should be noted that BinDNase predictions themselves are not sensitive to small changes in these parameters as the prediction performance of the BinDNase models are still higher than those of competing models. Finally, we wanted to evaluate that how well our supervised BinDNase method performs compared to unsupervised methods when BinDNase is trained on one cell type and applied to another cell type. Results for this comparison in Figure 9 show that BinDNase still outperforms PIQ even though the BinDNase model is learned from a different cell type. Overall, our results demonstrate that BinDNase is both accurate and generalizes to other cell types, thus making it a powerful and versatile tool.

3.7 Prediction accuracy saturates at a modest sequencing depth

The effect of sequencing depth on prediction accuracy was investigated by making predictions on the candidate binding sites using only subsamples of all reads (Fig. 11). For about half of the TFs for which the predictions are easy to make (i.e. AUC < 0.8) the required sequencing depth is much lower than the depth in the ENCODE DNase-seq data sets. Prediction accuracy saturates already at sequencing depth of about 30M–60 M, even though the original sequencing depth is as high as 270 M DNase-seq reads/samples. For TFs which are more difficult to predict (AUC 0.8 or below) sufficient saturation is achieved between 100 M and 150 M reads per sample. For a few TFs, such as EGR1 and RFX5, it seems that deeper sequencing could further improve the prediction results.

4 Discussion

Although transcription factor binding prediction using various high-throughput sequencing data has made significant progress recently, there is still an urgent demand to develop novel computational methods for analysing different kinds of sequencing data sets. This work sheds light on many questions considering computational analysis of deeply sequenced DNase I hypersensitivity data. Foremost, despite the fact that most previous methods have used traditional, low resolution canonical DNase I hotspot or footprint models, we demonstrate that for some TFs DNase-seq data contain high resolution information about TF-DNA interactions which can be used to improve discrimination between bound and unbound motif locations.

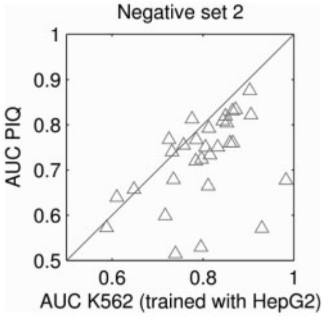

Fig. 9. The BinDNAse model trained on another cell type outperforms the unsupervised PIQ

The use of single nucleotide resolution DNase-seq data is hindered by the inherent sequence bias of the DNase molecule (He *et al.*, 2014) and therefore not all signals in the data discriminate bound and unbound sites. That is, it is not known *a priori* which part of the DNase-seq signal contain discriminative information. Thus, efficient feature selection methods, such as the one used in this work, are needed to construct accurate TF binding prediction models.

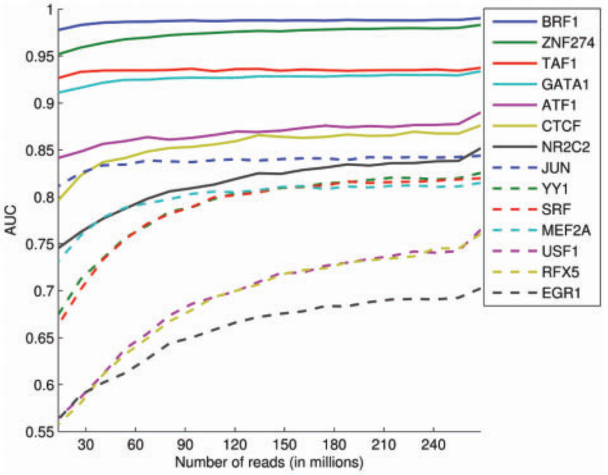

Fig. 11. Effect of sequencing depth on TF binding prediction accuracy. The AUC scores are computed for a representative collection of TFs using subsampled versions of the original DNase-seq data. The original sequencing depth is 270M reads

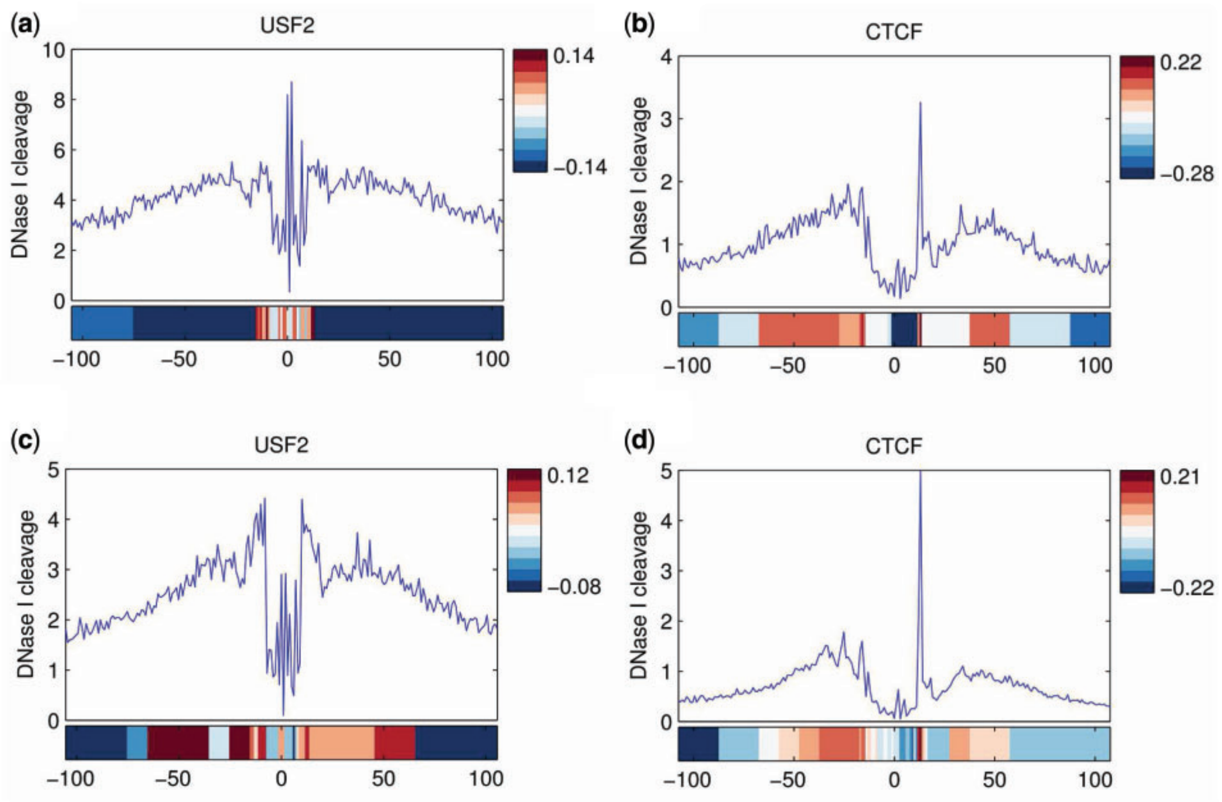

Fig. 10. The upper panels show the average DNase I cleavage centered at the TF binding motifs. The coloured bars indicate the optimised feature selection and their coefficients in the logistic regression model. Red (resp. blue) colour indicates positive (resp. negative) coefficient. The models trained with different cell types can be different (USF2) or highly similar (CTCF). **(a)** USF2 (HepG2), **(b)** CTCF (HepG2), **(c)** USF2 (K562) and **(d)** CTCF (K562)

This work also provides insight into which type of machine learning algorithms are suited for accurate TF binding predictions using DNase-seq data. According to our simulations supervised methods such as MILLIPEDE and BinDNase perform better than the commonly used unsupervised methods such as CENTIPEDE or PIQ. This is not an unexpected result because the supervised learning process contains an inbuilt mechanism for finding discriminative features as the class labels are known in the model training step. On the other hand, the unsupervised approach tries to find structure in the data without having any knowledge whether this structure truly contains discriminative power. The use of supervised model training explains at least partially why the algorithm performs well without any bias correction.

Although the supervised approach leads to a better prediction accuracy, estimating the models requires information about the true TF binding. Predictions can therefore be made only for transcription factors whose binding is measured with ChIP-seq or other techniques. The binding models however generalise to other cell types and the generalization implies the possibility of transferring ChIP-seq binding information to other cell types via the high resolution footprint models.

Our results do not imply superiority of the supervised methods for all prediction tasks. In some cases the accuracy of the unsupervised methods or even the simple DNase activity predictor is sufficient. The unsupervised methods are also useful when only a broad general view of the TF binding landscape is needed. The need for binding information for model building can be viewed as a disadvantage of the supervised models.

In this work we developed a novel method, BinDNase, for TF binding prediction using DNase-seq data. Via comprehensive simulations we show that BinDNase performs better than existing methods. We show that BinDNase's prediction accuracy is generally well-saturated at sequencing depths of the currently available DNase-seq data sets and that the method also generalises between cell types. We believe that BinDNase will be a useful tool in practise and help revealing the mechanisms of transcriptional regulation in numerous applications.

Acknowledgement

The authors also thank Tatsunori Hashimoto for his help with PIQ analysis.

Funding

This work was supported by the Academy of Finland Center of Excellence in Systems Immunology and Physiology and the Academy of Finland grant no. 255878.

Conflict of Interest: none declared.

References

Boyle,A.P. *et al.* (2008) High-resolution mapping and characterization of open chromatin across the genome. *Cell*, **132**, 311–322.

Buenrostro,J.D. *et al.* (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, dna-binding proteins and nucleosome position. *Nat. Methods*, **10**, 1213–1218.

Giresi,P.G. *et al.* (2006) Faire (formaldehyde-assisted isolation of regulatory elements) isolates active regulatory elements from human chromatin. *Genome Res.*, **17**, 877–885.

He,H.H. *et al.* (2014) Refined dnase-seq protocol and data analysis reveals intrinsic bias in transcription factor footprint identification. *Nat. Methods*, **11**, 73–78.

John,S. *et al.* (2011) Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.*, **43**, 264–268.

Kasowski,M. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.

Luo,K. and Hartemink,A.J. (2013) using dnase digestion data to accuractely identify transcription factor binding sites. *Pacific Symposium on Biocomputing*, 80–91.

Matsuda,M. *et al.* (1992) δ-thalassemia caused by disruption of the site for an erythroid-specific transcription factor, GATA-1, in the δ-globin gene promoter. *Blood*, **80**, 1347–1351.

Neph,S. *et al.* (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, **489**, 8390.

Piper,J. *et al.* (2013) Wellington: a novel method for the accurate identification of digital genomic footprints from dnase-seq data. *Nucleic Acids Res.*, **41**, e201.

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

Ramsey,S.A. *et al.* (2010) Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics*, **26**, 2071–2075.

Sherwood,R.I. *et al.* (2014) Discovery of directional and nondirectional pioneer transcription factors by modeling dnase profile magnitude and shape. *Nat. Biotechnol.*, **32**, 171–178.

Sung,M.-H. *et al.* (2014) Dnase footprint signatures are dictated by factor dynamics and dna sequence. *Mol. Cell*, **56**, 275–285.

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

Wang,J. *et al.* (2012) Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.*, **22**, 1798–1812.

Weirauch,M.T. *et al.* (2013) Evaluation of methods for modeling transcription factor sequence specificity. *Nat. Biotechnol.*, **31**, 126–134.

Wittwer,J. *et al.* (2006) Functional polymorphism in ALOX15 results in increased allele-spesific transcription in macrophages through binding of the transcription factor SPI1. *Hum. Mutat.*, **27**, 78–87.

Yardimci *et al.* (2014) Explicit DNAse sequence bias modeling enables high-resolution transcription factor footprint detection. *Nucleic Acids Res.*, **42**, 11865–11878.