Molecular BioSystems

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Mol. BioSyst.*, 2016, **12**, 371

Understanding the combinatorial action of transcription factors and microRNA regulation from regions of open chromatin†

Guantao Zheng,‡[a] Pan Zhang,‡[a] Zhihong Wu*[bc] and Dong Dong*[ab]

Received 20th October 2015,
Accepted 26th November 2015

DOI: 10.1039/c5mb00702j

www.rsc.org/molecularbiosystems

Transcriptional regulatory cascades are always triggered through the combinatorial interplay between transcription factors (TFs) and microRNAs (miRNAs) in eukaryotes. However, it is still a very substantial undertaking to dynamically profile their coordinated actions. In this work, we compared the differences in TFBS numbers between miRNA targets and non-targets, and found that miRNA targets tend to have more TFBS numbers. With the attempt to comprehensively understand the combinatorial action of TF and miRNA regulation from regions of open chromatin, we retrieved recently published DNase I hypersensitive sites (DHSs) across different human cell lines. The result showed that the differences are more statistically significant in DHS regions than non-DHS regions. Next, we trained classifiers for miRNA targets and non-targets. The result showed that TFBSs located in DHS regions achieved a competitive performance when discriminating miRNA targets and non-targets, whereas the performance of classifiers using TFBSs located in non-DHS regions is close to that of a random classifier. After the DHSs were divided into intergenic, transcription start sites (TSSs) and gene body DHS regions based on their genomic locations, only TFBSs located in TSS DHS regions provided a competitive performance. Our results provide us a clue that the coordinated activity of miRNAs and TFs describing the mechanism of gene expression control should be examined in a dynamic perspective.

Introduction

The transcriptional regulation of gene expression plays a pivotal role in the processes of differentiation, proliferation and development in the eukaryotic system.[1] The underlying molecular mechanisms of transcriptional regulation are multilayered and provide a complicated system to decipher.[1,2] Both miRNAs and TFs have been considered as two major classes of *trans*-regulators, and play an important role in transcriptional regulation. To date, many studies have demonstrated the coordinated action of TFs and miRNAs exerted on gene expression control.[3–6]

Systematic computational approaches have an advantage in elucidating coordination between miRNA and TF regulation. Recently, Cui *et al.* presented that genes with more TF binding sites (TFBSs) are likely to be targeted by miRNAs,[3] and it can be elucidated that a tight relationship exists between the complex regulatory network at the transcription level and that controlled by miRNAs at the post-transcriptional level. Moreover, it has been documented that miRNAs co-evolved with TFs, and the rapidly evolved TFs preferentially activate miRNAs.[7] The coordinated activity of miRNAs and TFs described the complicated mechanisms of transcriptional control.

The spatiotemporal pattern of gene expression is determined by a dynamic transcriptional machinery, however, previous studies only showed a fairly static 'snap-shot' of transcriptional regulation based on the computational approaches. For example, previous sequence-based approaches generated myriads of putative TF binding sites (TFBSs), however, only a small fraction is functional in a particular context.[8] With the scale-up of the ENCODE project,[9] an increasing amount of data is becoming available to describe the molecular mechanism of transcription. As we know, the function of *trans*-regulators involved in gene expression is controlled by the dynamic organization of chromatin inasmuch it defines the chromatin accessibility.[10] The induction of a more 'open' chromatin region can increase the accessibility of *trans*-regulators to the genomic DNA, which is an important factor for the regulation of gene expression. The chromatin accessible regions are organized

[a] *Laboratory of Molecular Ecology and Evolution, Institute of Estuarine and Coastal Research, East China Normal University, Shanghai, 200062, P. R. China*
[b] *Beijing Key Laboratory for Genetic Research of Bone and Joint Disease, No. 1 Shuaifuyuan, Beijing, 100730, P. R. China. E-mail: ddong.ecnu@gmail.com, wuzh3000@126.com*
[c] *Central laboratory, Peking Union Medical College Hospital, Peking Union Medical College and Chinese Academy of Medical Sciences, No. 1 Shuaifuyuan, Beijing, 100730, P. R. China*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c5mb00702j

‡ These authors contributed equally to this work.

with dozens to hundreds of co-activated elements. So, chromatin accessibility data provided an alternative measure of cooperative binding of *trans*-acting factors in place of a canonical nucleosome.[11–14] DNase I hypersensitive sites (DHSs) are short chromatin regions that are highly sensitive to cleavage by DNase I enzyme. Most of the DHSs typically occur in nucleosome-depleted regions and arise as a result of TF binding. DNase-seq data not only provide the locations of regulatory elements, but also characterize a complete chromatin landscape.[14] Recent studies profiling open chromatins in a genome-wide fashion[14,15] open up the door for us to dynamically explore the active TFBS across different human cell lines.

In this work, we retrieved recently published DNase-seq data across diverse human cell lines representing a wide variety of tissues, and these data show a clear advantage in representing active TFBSs. Because of the importance of the concerted action of miRNA and TF regulation, we attempted to dynamically analyze the actions between miRNAs and TF regulation based on the chromatin accessibility landscape. Our work provided a guide to deeply understand the complexity of gene regulation.

Materials and methods

Gene annotation

Human protein-coding genes were retrieved from the Ensembl database (Release 71).[16] The genes less than 2 kb long were removed to avoid the inclusion of non-genic regions downstream of a gene. We also removed all genes having a promoter within 2 kb of another gene to avoid analyzing upstream regions overlapping with another gene.

DNase-seq data and ChIP-seq data

In this work, we downloaded the DNase-seq data of 19 human cell lines[14,15] from the UCSC genome browser (http://genome.ucsc.edu/), representing a wide variety of human tissues. The sequencing raw data were aligned to the human reference genome (built hg.19) using BWA.[17] F-seq, a kernel density estimator, was used for peak calling with default parameters.[18] The identified peak regions were considered as DHSs or open chromatin regions. A total of 539 175 distinct high-confidence DHSs were obtained and each of them was active in one or more cell types. Next, we downloaded previously identified DHSs in 19 cell lines[14,15] from the UCSC genome browser, and a total of 480 626 distinct DHSs were found, among which 476 360 DHSs were also identified in our results. As for the TFBSs, we totally downloaded 229 ChIP-seq datasets[19] from the UCSC genome browser, representing the DNA footprints of 59, 120, 86 and 64 TFs in four human cell lines (H1, K562, GM12878 and HepG2), respectively. A total of 282 982, 689 191, 593 813 and 589 960 ChIP-seq peaks were found in H1, K562, GM12878 and HepG2 cell lines, respectively.

DHSs were classified according to the genomic regions of genes. If a DHS is located in the transcription start site region (TSS) of any transcript isoforms of a gene, it was classified as a TSS DHS for the focal gene. Those DHSs were classified as gene body DHSs if they overlap with any regions of the exons or introns, and all other DHSs which do not overlap with any region of a gene were classified as intergenic DHSs. For each intergenic DHS, we used BEDTOOLS software[20] to find the nearest gene, and associated it with that DHS if the distance between them was less than 200 kb.

Annotation of miRNA targets and non-targets

The miRNA targets were taken from three previously published *in silico* miRNA target prediction methods, including TargetScan (http://www.targetscan.org version 5.1),[21] PITA (http://genie.weizmann.ac.il/pubs/mir07/mir07_data.html)[22] and Pictar (http://genome.ucsc.edu four-way).[23] Targets predicted by TargetScan with a total context score < -0.3 were removed, and those targets with at least one conserved 7-mer or 8-mer were chosen as reliable miRNA targets. For PITA targets, a score less than -10 was selected as the threshold to choose reliable miRNA targets. To minimize the false positive of miRNA target prediction, a high-quality miRNA target data set was generated by intersecting data generated by at least two different *in silico* miRNA target prediction methods. Those without being detected by any methods were defined as miRNA non-targets. At last, a total of 5298 genes were identified to be miRNA targets, and 10 203 genes were identified to be miRNA non-targets, respectively. The average number of targets per miRNA is 170.

Position weight matrix (PWM) scan

We compiled PWMs for vertebrate TFs from the JASPAR,[24] TRANSFAC[25] and Uniprobe databases,[26] which gave us a total of 789 PWMs. The ''PWM Score'' is a log-likelihood ratio of the probability of a given sequence under the PWM model, compared to a random sequence model. Each TF is represented by a specific PWM (a matrix of frequencies) with which this TF is expected to bind a certain DNA motif. Each PWM was used to score the intergenic, TSS and gene body DHS regions while looking for subsequences that closely match the binding motif represented by the PWM. Next, we scanned the sequence from each DHS and non-DHS region. For each location, a score was calculated based on the probability that the sequence was generated in the PWM model *versus* the probability that the specific sequence was generated in the background model. The first-order Markov Model trained on a 500-bp window centered at the base pair was applied in the background model. This method could effectively correct for the underlying dinucleotide composition and separate the signal from noise.[27]

The scores were generated for each base pair. A 60-bp sliding window was moved across the sequence, and we summed scores of all base pairs in each window. The maximum window score was determined as the TFBS score for that TF. This sliding window based method could account for local clustering of binding sites, which have been shown to be more likely to be bound by TFs than single binding sites. In general, one gene may be associated with more than one DHS, and these regions were assumed to be the putative regulatory regions for that gene. Here, we assigned the maximum TFBS score of that region to each gene.

Support vector machine

Based on the miRNA–mRNA relationship, genes can be classified into miRNA targets and non-targets. TFBS scores were used as features for SVM classifiers to discriminate miRNA targets and non-targets. All the features were integrated using the SVM model. We employed radial kernel for training and predicting in the SVM classification model. Next, we evaluated the performance of the models using five-fold cross-validation. We randomly divided the data into five subsets with equal sizes, four training sets and one testing set, respectively. The model was trained using the training set and applied to the testing set to predict expression. The prediction power of the SVM model was estimated based on the testing set. The model generates a probability indicating how likely a gene tends to be targeted by miRNAs. By setting different threshold values, we can depict the sensitivity (true positive rate) and the specificity (true negative rate) of the prediction. The receiver operator characteristic (ROC) curve was used to show the classification accuracy of our SVM model. In this work, AUCs (area under the curve) were calculated from each cell line in the intergenic DHS region, the TSS DHS region, the gene body DHS region and non-DHS regions, respectively.

Based on the above-described method, we constructed models for classifying miRNA targets and non-targets, with each gene having 789 features. A support vector machine (SVM), implemented by the LibSVM package,[28] was introduced for classification.

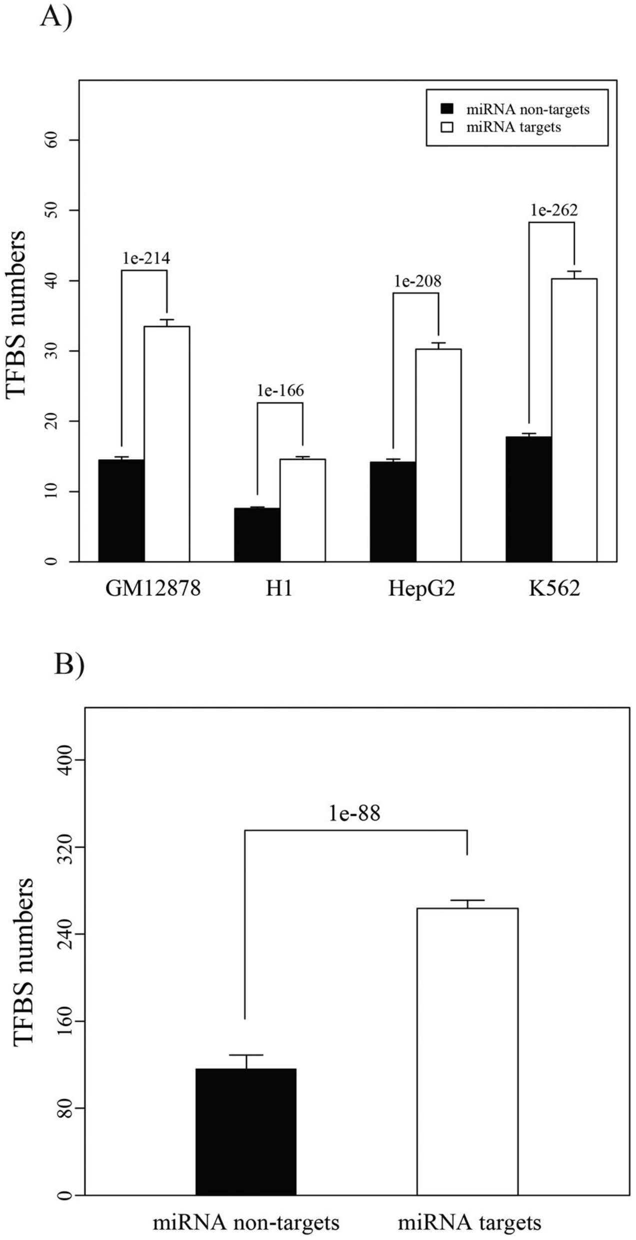

Fig. 1 Differential number of TFBSs between miRNA targets and non-targets. (A) miRNA targets have a significantly higher number of TFBSs generated by the ChIP-seq method than non-targets across 4 human cell lines. (B) miRNA targets have a significantly higher number of TFBSs predicted by PWMs than non-targets. The error bars represent the 95% confidence intervals.

Results

Combinatorial regulation of TF and miRNA at the open chromatin regions

Since TFs and miRNA are principal classes of gene regulators, their combinatorial regulation of TF and miRNA has attracted extensive attention. In order to dynamically measure the co-regulation of TFs and miRNAs, ChIP-seq data in human four cell lines were used. In line with previous work,[3] our results showed that genes with more TFBSs have a higher probability to be targeted by miRNAs ($P = 1 \times 10^{-214}$ for GM12878 cell, $P = 1 \times 10^{-166}$ for H1 cell, $P = 1 \times 10^{-208}$ for HepG2 cell, $P = 1 \times 10^{-262}$ for K562 cell, the Wilcoxon rank-sum test, Fig. 1A). In order to explore whether our results are sensitive to the way of miRNA target prediction, we defined miRNA targets in four different ways and then re-analyzed the difference of TFBS numbers between miRNA targets and non-targets. We selected targetScan prediction, PITA prediction and picta prediction to define miRNA targets, and also collected experimentally verified miRNA target data from the miRTarBase database, respectively. The result showed similar patterns that genes with more TFBSs have a higher probability to be targeted by miRNAs, no matter which miRNA target definition was selected (Fig. S1, ESI†). Next, we randomly defined the dataset of miRNA targets and non-targets, and we compared the differences in TFBS numbers between these two datasets in four human cell lines (GM12878, H1, HepG2 and K562). This step was repeated 1000 times. Then, we plotted the distribution of these *P*-values and found that most of them were larger than 0.05 (Fig. S2, ESI†), which indicated that the difference of TFBS numbers between miRNA targets and non-targets is not a random event. The ChIP-seq method has become the gold standard method for the genome-wide detection of the binding locations for individual TFs, however, ChIP-seq data are still limited and the DNA footprints of many TFs have not been completely measured. Next, we scanned the genome for all TFBSs with substantial similarity to position weight matrices (PWMs) from TRANSFAC, JASPAR, and UniProbe databases. We obtained a similar result that miRNA targets tend to have more

Table 1 Comparison of TFBS numbers between miRNA targets and non-targets in DHS regions and non-DHS regions, respectively. The numbers represent TFBS numbers of each gene in each DHS or non-DHS region

	DHS regions			Non-DHS regions		
Cell lines	miRNA targets	miRNA non-targets	*P*-value	miRNA targets	miRNA non-targets	*P*-value
AosmcSerumfree	0.022	0.008	1.01×10^{-312}	0.002	0.0008	1.06×10^{-89}
Chorion	0.015	0.006	1.08×10^{-288}	0.0018	0.001	9.43×10^{-92}
Gliobla	0.022	0.009	1.66×10^{-298}	0.0016	0.0008	3.26×10^{-84}
GM12878	0.015	0.006	1.23×10^{-288}	0.0017	0.0009	7.98×10^{-82}
H1	0.01	0.005	4.62×10^{-265}	0.0018	0.0009	5.97×10^{-100}
Helas3	0.019	0.008	1.75×10^{-302}	0.002	0.001	2.09×10^{-92}
Hepatocytes	0.014	0.006	1.91×10^{-277}	0.0024	0.001	7.58×10^{-129}
HepG2	0.017	0.008	8.55×10^{-283}	0.0017	0.0008	1.65×10^{-75}
Hmec	0.021	0.009	1.44×10^{-301}	0.0023	0.001	3.72×10^{-78}
Hsmmt	0.018	0.008	2.87×10^{-270}	0.0023	0.001	1.68×10^{-84}
Huvec	0.019	0.007	4.11×10^{-269}	0.0017	0.0009	3.98×10^{-75}
K562	0.015	0.007	6.68×10^{-244}	0.0018	0.001	5.69×10^{-71}
Lncap	0.012	0.006	1.23×10^{-232}	0.002	0.001	5.37×10^{-85}
Mcf7	0.018	0.008	3.56×10^{-279}	0.0017	0.0009	1.47×10^{-77}
Medullo	0.011	0.005	4.28×10^{-222}	0.0021	0.001	1.04×10^{-83}
Melano	0.019	0.007	7.11×10^{-299}	0.0017	0.0009	4.11×10^{-68}
Nhek	0.023	0.009	2.16×10^{-342}	0.002	0.001	4.12×10^{-81}
Osteobl	0.017	0.007	9.18×10^{-284}	0.002	0.001	2.81×10^{-85}
Progfib	0.02	0.008	3.11×10^{-305}	0.0018	0.0009	1.41×10^{-83}

TFBSs than non-targets ($P = 1 \times 10^{-88}$, the Wilcoxon rank-sum test, Fig. 1B). The computational methods search over the genome for motifs representing the DNA binding sequence for TFs and generated myriads of putative results. However, only a small fraction of these putative TFBSs is functional in a particular context. Since TFBSs are concentrated in open chromatin regions and DHS regions are well correlated with TFBSs,[13,14] genome-wide evidence shows that TF binding to DNA regions is a key driver of the chromatin accessibility. In this work, those TFBSs located in DHS regions were regarded as active transcriptional regulatory elements. We retrieved DNase-seq data in 19 human cell lines, which were benefitted from extensive experiments performed by different ENCODE production groups. The differential accessible patterns between miRNA targets and non-targets were investigated across 19 various human cell lines. As shown in Table 1, the differences in TFBS numbers between miRNA targets and non-targets are more statistically significant in DHS regions than non-DHS regions in each cell line. The detection of active TFBSs with the combination of DNase I hypersensitivity provided a dynamic scenario to illustrate the functional TF binding in a particular context. Our results broaden our understanding of the combinatorial regulation of TFs and miRNAs in a dynamic perspective.

Classifying miRNA targets and non-targets from sequence features in open chromatin

Next, we used DHS data of 19 human cell lines to determine whether miRNA targets and non-targets can be distinguished using regulatory regions with open chromatin. We compiled position weight matrices (PWM) for TFs from TRANSFAC, JASPAR and UniProbe databases. TFBS scores which account for local dinucleotide compositions were calculated based on these PWMs. Briefly, TFBS scores were assigned to the locations in the sequence based on the probability of the PWM generating the specific sequence *versus* the probability that the sequence was generated by a background model (see Materials and methods). A total of 789 PWMs were used to measure TFBS scores that accounted for local dinucleotide composition. For each DHS, we assigned the maximum sliding window score as a TFBS score. A gene may be associated with more than one DHS, and we took the maximum TFBS score across all associated DHSs.

TFBS scores were used as features for SVM classifiers to discriminate between miRNA targets and non-targets. At first, we built SVM classifiers on the task to discern whether a gene is the target of miRNAs. The performance of the model was assessed by five-fold cross-validations. AUCs were used to evaluate the performance of a model, where higher AUC values represent more perfect classification. The result showed that the AUCs vary slightly across diverse human cell lines, and it is obvious that the performance of the classifier using non-DHS region information (median AUCs = 0.59) is close to that of a random classifier, whereas the performance of the classifier using DHS region information (median AUCs = 0.73) is significantly higher than that of the non-DHS region (Fig. 2A). Next, we classified DHSs into intergenic DHSs, TSS DHSs and gene body DHSs based on their genomic locations (Fig. 2B). Then, SVM classifiers in intergenic DHS regions, TSS DHS regions, gene body DHS regions were built, respectively. The performance of the classifiers in intergenic DHSs (median AUCs = 0.53) and gene body DHSs (median AUCs = 0.6) is close to that of non-DHS regions (median AUCs = 0.59) (Fig. 2A). However, the classifiers using TSS DHS sequences display stronger improvements in performance (median AUCs = 0.76) compared with intergenic DHSs and gene body DHSs. It suggests that strong performance of the classifier is achieved by scanning for TFBS matches in the accessible chromatin at the TSS regions, while no obvious evidence supports the coordinated action between miRNAs and those TFs binding to intergenic and gene body DHSs. To further validate our results, we separated TFBSs in the human K562 cell line generated by the ChIP-seq method into intergenic DHS, TSS DHS and gene body DHS regions, respectively. SVM classifiers were built by using ChIP-seq tag density

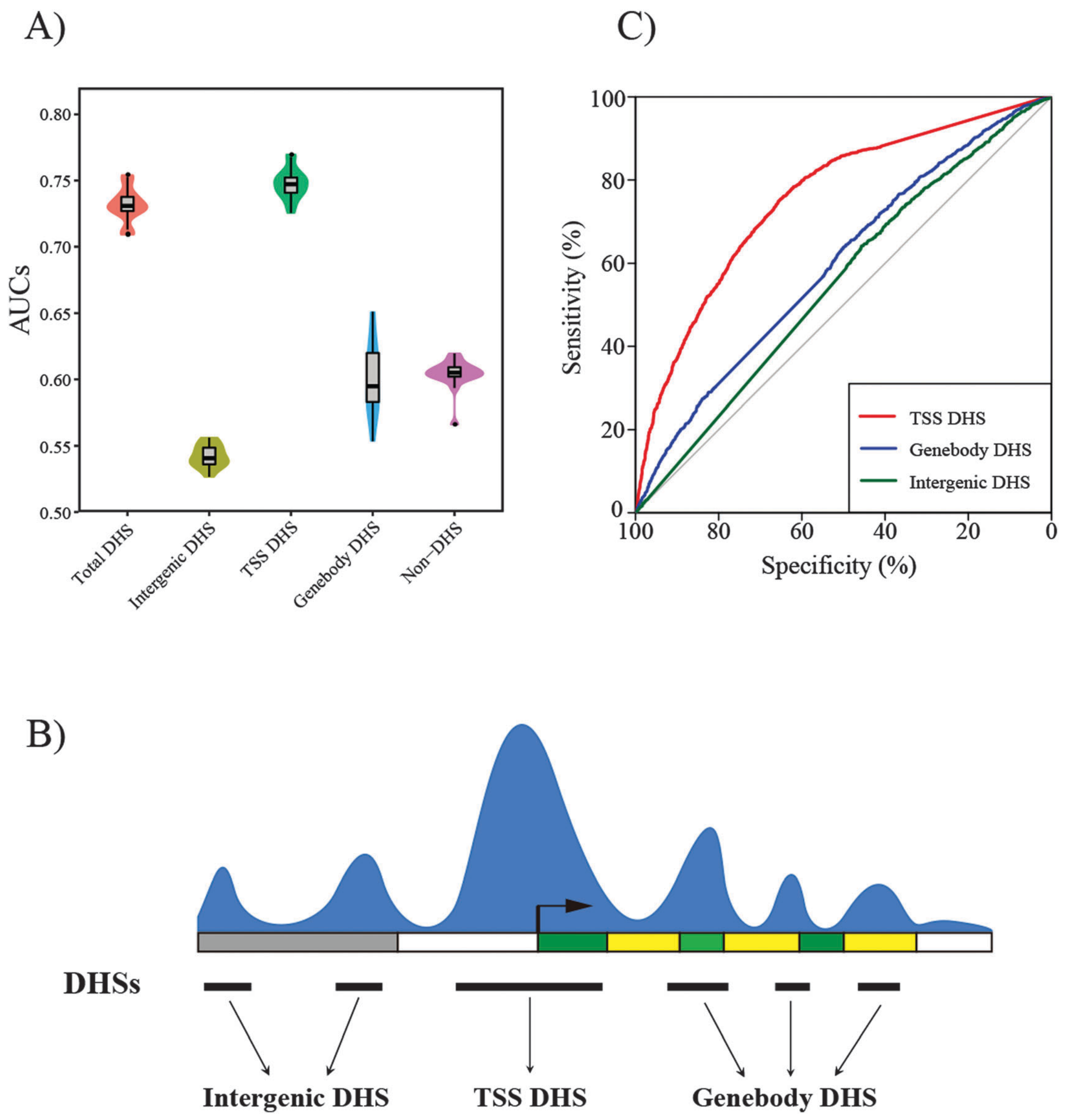

Fig. 2 The performance of classifiers for miRNA targets and non-targets. (A) Violin plot showing that the classifiers using TSS DHSs display stronger improvements in performance (median AUCs ~ 0.76) compared with intergenic DHS and gene body DHS and non-DHS regions; (B) properties of DHSs based on genomic regions. DHSs are intergenic and those that are overlapping with TSS and gene body were classified into TSS and gene body DHS, respectively; (C) ROC curves showing the performance of classifiers based on whole-genome wide ChIP-seq data of TFs in the K562 cell line at the intergenic, TSS and gene body DHS regions, respectively.

as a feature. A similar result showed that only TFBSs located in TSS DHS regions display strong improvements of performance (AUC = 0.77) (Fig. 2C). All results clearly indicated that the coordinated activities between TF binding and miRNA regulation are dynamic processes and mainly confined to those TFs binding in the open chromatin around TSS regions. Lessons should be learned when globally detecting the concerted regulation of TFs and miRNAs.

Prediction of miRNA–mRNA targeting relationships and functional implications of coordinated regulators

Since we have proved that TFBSs located in TSS DHS regions display stronger performance in miRNA target prediction, we wonder whether TFBS scores in TSS DHS regions are for miRNA–mRNA targeting relationship. SVM classifiers were trained to discriminate targets and non-targets for those miRNAs having at least 30 targets in the positive golden standard, and this procedure resulted in 226 miRNAs for which the predictions were performed. In Fig. 3, the results showed the AUC scores for all 226 miRNAs as a function of the size of the training set, and the average AUCs were estimated at 79.4% in the human K562 cell line. The accuracy is still high with the increasing of the number of positives.

Discussion

Transcription is a complicated dynamic process, involving a combination of various regulators.[29] In this study, we undertook

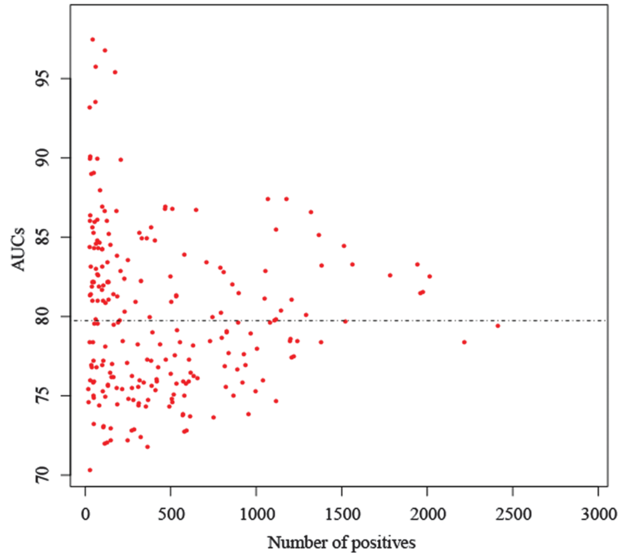

Fig. 3 miRNA targets prediction accuracy. Area under the ROC curve as a function of the number of positive examples for miRNAs. Dotted line indicates the average case.

a comprehensive analysis of the relationship between miRNAs and TFs based on chromatin accessibility data across diverse human cell lines. We found that the combinatorial action of miRNA and TF regulation is general, and it is especially pronounced at the open chromatin in the TSS regions. Our work indicated a complicated interconnection between miRNAs and TF regulation machinery, which could provide a useful starting point to explore the dynamic molecular basis of transcriptional regulation complexity.

The interactions of *trans*-acting factors and DNA targets involved in gene expression are controlled by the dynamic organization of chromatin inasmuch as it defines the chromatin accessibility.[30] Chromatin accessibility using DNase-seq provides a powerful way to link genome sequence changes to gene expression variations.[11,31] DHSs have some properties that make them especially valuable for the delineation of the genomic *cis*-regulatory compartment. First, unlike other strategies that require multiple ChIP-seq data for highly informative *trans*-acting factors, DNase-seq identifies most regions of the genome that are accessible to *trans*-acting factor binding. Second, DHSs are central to all defined classes of active *cis*-regulatory elements including promoters, enhancers, insulators and locus control regions, which reveal all *cis*-components of transcription through a single genome-wide assay. Third, DHSs represent ~2% of the human genome, which makes them easy to focus on DNA regions involved in transcriptional regulation. Lastly, chromatin is a dynamic structure that responds to myriad stimuli to regulate access to DNA. As part of the ENCODE project, DNase-seq data provide us opportunity to delineate the dynamic nature of chromatin and TF binding properties.

Recently, it has been observed that a relationship exists between gene regulation networks controlled by TFs at the transcriptional level and those controlled by miRNAs at the post-transcriptional level.[3,32] Our work has the following advantages to help us understand the combinatorial action of miRNAs and TF. First, previous studies showed a static 'snap-shot' of the active concert of these regulators based on motif-predicted binding sites. Here, we presented the coordinated activity from a broad and dynamic perspective since DHSs are central to all defined classes of active *cis*-regulatory elements. Second, the coordinated activities between TF binding and miRNA regulation are confined to those TFs binding at the open chromatin region around TSS regions, instead of those TFs binding at intergenic and gene body regions. As we know, TF binding is always concentrated in a highly accessible chromatin structure that shows a nucleosome-depleted organization. TF binding to DNA is a dynamic process, and the complex spatio-temporal regulations of gene expression are thought to be achieved by the complex transcriptional machinery.[33] Our results indicated that the coordinated activity model may be applied to a relatively broaden dataset of TFBSs based on the evidence presented here in various cell lines.

Since gene expression can be influenced by both TFs and miRNAs, the effect of the combinatorial regulation of miRNA and TF might be learned from gene expression variation. Here, we also trained classifiers for different expression patterns using TF binding information in TSS DHS regions in 19 human cell lines. In line with a previous work,[15] the competitive performance was obtained when classifying cell-line-specific expressed genes and constitutively expressed genes. Recently, the impact of miRNA regulation on variations in human gene expression has been widely examined.[34] We found that miRNA targets are highly enriched in cell-line-specific expressed genes and have higher expression variation among human cell lines. This finding suggested that the coordinated regulation of TFs and miRNAs is a general property in mRNA transcription.

We should note that the approach of predicting miRNA targets from the *cis*-regulatory sequence as presented here is impeded by several limitations. First, only a small fraction of TFBSs are known, which restricts the application of our model. Another challenge is mapping *cis*-regulatory elements to the genes they regulate. Here, we simply assumed that the nearest gene is the most likely target. Accounting for long-range regulatory interactions by methods like 3C, 4C, 5C and Hi–C allows for more accurately connecting DHSs to their correct target genes.[35] Our work remains rudimentary, and further efforts are needed to establish more informative connections between miRNAs and chromatin features. False positives notwithstanding, the results provided valuable staring points for follow-up studies.

Taken together, our work presented the combinatorial regulation of TF and miRNAs in a dynamic perspective. We speculated that the emerging pictures of transcription regulation are much more complicated than previously thought. This study comprehensively provided the attempt to understand the complexity of gene regulation control.

Conflict of interests

The authors declare that they have no competing interests.

Authors' contributions

DD designed the study, and GZ, PZ and DD carried out the data analysis. DD wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This work is supported by National Natural Science Foundation of China to Dong Dong (Grant No. is 31200956).

References

1 H. Yu, N. M. Luscombe, J. Qian and M. Gerstein, *Trends Genet.*, 2003, **19**, 422–427.

2 T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford and R. A. Young, *Science*, 2002, **298**, 799–804.

3 Q. Cui, Z. Yu, Y. Pan, E. O. Purisima and E. Wang, *Biochem. Biophys. Res. Commun.*, 2007, **352**, 733–738.

4 C. Y. Chen, S. T. Chen, C. S. Fuh, H. F. Juan and H. C. Huang, *BMC Bioinf.*, 2011, **12**(suppl 1), S41.

5 A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander and D. S. Marks, *Genome Biol.*, 2003, **5**, R1.

6 A. Marson, S. S. Levine, M. F. Cole, G. M. Frampton, T. Brambrink, S. Johnstone, M. G. Guenther, W. K. Johnston, M. Wernig, J. Newman, J. M. Calabrese, L. M. Dennis, T. L. Volkert, S. Gupta, J. Love, N. Hannett, P. A. Sharp, D. P. Bartel, R. Jaenisch and R. A. Young, *Cell*, 2008, **134**, 521–533.

7 K. Chen and N. Rajewsky, *Nat. Rev. Genet.*, 2007, **8**, 93–103.

8 G. A. Maston, S. K. Evans and M. R. Green, *Annu. Rev. Genomics Hum. Genet.*, 2006, **7**, 29–59.

9 N. de Souza, *Nat. Methods*, 2012, **9**, 1046.

10 D. S. Gross and W. T. Garrard, *Annu. Rev. Biochem.*, 1988, **57**, 159–197.

11 A. P. Boyle, L. Song, B. K. Lee, D. London, D. Keefe, E. Birney, V. R. Iyer, G. E. Crawford and T. S. Furey, *Genome Res.*, 2011, **21**, 456–464.

12 R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad and J. K. Pritchard, *Genome Res.*, 2011, **21**, 447–455.

13 S. Neph, J. Vierstra, A. B. Stergachis, A. P. Reynolds, E. Haugen, B. Vernot, R. E. Thurman, S. John, R. Sandstrom, A. K. Johnson, M. T. Maurano, R. Humbert, E. Rynes, H. Wang, S. Vong, K. Lee, D. Bates, M. Diegel, V. Roach, D. Dunn, J. Neri, A. Schafer, R. S. Hansen, T. Kutyavin, E. Giste, M. Weaver, T. Canfield, P. Sabo, M. Zhang, G. Balasundaram, R. Byron, M. J. MacCoss, J. M. Akey, M. A. Bender, M. Groudine, R. Kaul and J. A. Stamatoyannopoulos, *Nature*, 2012, **489**, 83–90.

14 R. E. Thurman, E. Rynes, R. Humbert, J. Vierstra, M. T. Maurano, E. Haugen, N. C. Sheffield, A. B. Stergachis, H. Wang, B. Vernot, K. Garg, S. John, R. Sandstrom, D. Bates, L. Boatman, T. K. Canfield, M. Diegel, D. Dunn, A. K. Ebersol, T. Frum, E. Giste, A. K. Johnson, E. M. Johnson, T. Kutyavin, B. Lajoie, B. K. Lee, K. Lee, D. London, D. Lotakis, S. Neph, F. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, A. Safi, M. E. Sanchez, A. Sanyal, A. Shafer, J. M. Simon, L. Song, S. Vong, M. Weaver, Y. Yan, Z. Zhang, B. Lenhard, M. Tewari, M. O. Dorschner, R. S. Hansen, P. A. Navas, G. Stamatoyannopoulos, V. R. Iyer, J. D. Lieb, S. R. Sunyaev, J. M. Akey, P. J. Sabo, R. Kaul, T. S. Furey, J. Dekker, G. E. Crawford and J. A. Stamatoyannopoulos, *Nature*, 2012, **489**, 75–82.

15 A. Natarajan, G. G. Yardimci, N. C. Sheffield, G. E. Crawford and U. Ohler, *Genome Res.*, 2012, **22**, 1711–1722.

16 P. Flicek, I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, C. Garcia-Giron, L. Gordon, T. Hourlier, S. Hunt, T. Juettemann, A. K. Kahari, S. Keenan, M. Komorowska, E. Kulesha, I. Longden, T. Maurel, W. M. McLaren, M. Muffato, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, G. R. Ritchie, M. Ruffier, M. Schuster, D. Sheppard, D. Sobral, K. Taylor, A. Thormann, S. Trevanion, S. White, S. P. Wilder, B. L. Aken, E. Birney, F. Cunningham, I. Dunham, J. Harrow, J. Herrero, T. J. Hubbard, N. Johnson, R. Kinsella, A. Parker, G. Spudich, A. Yates, A. Zadissa and S. M. Searle, *Nucleic Acids Res.*, 2013, **41**, D48–D55.

17 H. Li and R. Durbin, *Bioinformatics*, 2010, **26**, 589–595.

18 A. P. Boyle, J. Guinney, G. E. Crawford and T. S. Furey, *Bioinformatics*, 2008, **24**, 2537–2538.

19 M. B. Gerstein, A. Kundaje, M. Hariharan, S. G. Landt, K. K. Yan, C. Cheng, X. J. Mu, E. Khurana, J. Rozowsky, R. Alexander, R. Min, P. Alves, A. Abyzov, N. Addleman, N. Bhardwaj, A. P. Boyle, P. Cayting, A. Charos, D. Z. Chen, Y. Cheng, D. Clarke, C. Eastman, G. Euskirchen, S. Frietze, Y. Fu, J. Gertz, F. Grubert, A. Harmanci, P. Jain, M. Kasowski, P. Lacroute, J. Leng, J. Lian, H. Monahan, H. O'Geen, Z. Ouyang, E. C. Partridge, D. Patacsil, F. Pauli, D. Raha, L. Ramirez, T. E. Reddy, B. Reed, M. Shi, T. Slifer, J. Wang, L. Wu, X. Yang, K. Y. Yip, G. Zilberman-Schapira, S. Batzoglou, A. Sidow, P. J. Farnham, R. M. Myers, S. M. Weissman and M. Snyder, *Nature*, 2012, **489**, 91–100.

20 A. R. Quinlan and I. M. Hall, *Bioinformatics*, 2010, **26**, 841–842.

21 B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel and C. B. Burge, *Cell*, 2003, **115**, 787–798.

22 M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul and E. Segal, *Nat. Genet.*, 2007, **39**, 1278–1284.

23 A. Krek, D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel and N. Rajewsky, *Nat. Genet.*, 2005, **37**, 495–500.

24 J. C. Bryne, E. Valen, M. H. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard and A. Sandelin, *Nucleic Acids Res.*, 2008, **36**, D102–D106.

25 E. Wingender, X. Chen, R. Hehl, H. Karas, I. Liebich, V. Matys, T. Meinhardt, M. Pruss, I. Reuter and F. Schacherer, *Nucleic Acids Res.*, 2000, **28**, 316–319.

26 K. Robasky and M. L. Bulyk, *Nucleic Acids Res.*, 2011, **39**, D124–D128.

27 M. Megraw, F. Pereira, S. T. Jensen, U. Ohler and A. G. Hatzigeorgiou, *Genome Res.*, 2009, **19**, 644–656.

28 C. C. Chang and C. J. Lin, *LIBSVM: a library for support vector machines*, 2001, available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

29 J. Yu, M. A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J. L. Frane, S. Tian, J. Nie, G. A. Jonsdottir, V. Ruotti, R. Stewart, Slukvin II and J. A. Thomson, *Science*, 2007, **318**, 1917–1920.

30 A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey and G. E. Crawford, *Cell*, 2008, **132**, 311–322.

31 J. R. Hesselberth, X. Chen, Z. Zhang, P. J. Sabo, R. Sandstrom, A. P. Reynolds, R. E. Thurman, S. Neph, M. S. Kuehn, W. S. Noble, S. Fields and J. A. Stamatoyannopoulos, *Nat. Methods*, 2009, **6**, 283–289.

32 S. Arora, R. Rana, A. Chhabra, A. Jaiswal and V. Rani, *Mol. Genet. Genomics*, 2013, **288**, 77–87.

33 S. E. McGuire, G. Roman and R. L. Davis, *Trends Genet.*, 2004, **20**, 384–391.

34 R. Zhang and B. Su, *Nucleic Acids Res.*, 2008, **36**, 4621–4628.

35 B. van Steensel and J. Dekker, *Nat. Biotechnol.*, 2010, **28**, 1089–1095.