BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 2 2008, pages 165–171
doi:10.1093/bioinformatics/btm572

Sequence analysis

Identification of phylogenetically conserved microRNA *cis*-regulatory elements across 12 *Drosophila* species

Xiaowo Wang[1], Jin Gu[1], Michael Q. Zhang[1,2] and Yanda Li[1,*]

[1]MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TNLIST/Department of Automation, Tsinghua University, Beijing 100084, China and [2]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

Received on May 14, 2007; revised on July 18, 2007; accepted on November 9, 2007

Advance Access publication November 24, 2007

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: MicroRNAs are a class of endogenous small RNAs that play regulatory roles. Intergenic miRNAs are believed to be transcribed independently, but the transcriptional control of these crucial regulators is still poorly understood.

Results: In this work, phylogenetic footprinting is used to identify conserved *cis*-regulatory elements (CCEs) surrounding intergenic miRNAs in Drosophila. With a two-step strategy that takes advantage of both alignment-based and motif-based methods, we identified CCEs that are conserved across the 12 fly species. When compared with TRANSFAC database, these CCEs are significantly enriched in known transcription factor binding sites (TFBSs). Moreover, several TFs that play essential roles in *Drosophila* development (e.g. *Adf-1*, *Abd-B*, *Sd*, *Prd*, *Ubx*, *Zen* and *En*) are found to be preferentially regulating the miRNA genes. Further analysis revealed many over-represented *cis*-regulatory modules (CRMs) composed of multiple known TFBSs, motif pairs with significant distance constraints and a number of novel motifs, many of which preferentially occur near the transcription start site of protein-coding genes. Additionally, a number of putative miRNA-TF regulatory feedback loops were also detected.

Availability: Supplementary Material and the Perl scripts performing two-step phylogenetic footprinting are available at http://bioinfo.au.tsinghua.edu.cn/member/xwwang/mircisreg

Contact: daulyd@tsinghua.edu.cn

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

1 INTRODUCTION

MicroRNAs (miRNAs) are a class of ~22 nt long endogenous small RNA molecules that play essential regulatory roles in diverse organisms (Bartel, 2004). In animal cells, intergenic miRNAs are generally transcribed by RNA polymerase II (Cai *et al.*, 2004; Lee *et al.*, 2004; Zhou *et al.*, 2007), although some by RNA polymerase III (Borchert *et al.*, 2006). The long primary RNA transcripts (also called pri-miRNAs) (Cai *et al.*, 2004) are subsequently processed through a two-step process to produce ~70 nt hairpin-like precursors (pre-miRNAs) and ~22 nt mature miRNAs, by two RNase III enzymes Drosha and Dicer (Bartel, 2004), respectively. These tiny RNA molecules can direct the posttranscriptional regulation of target mRNAs for degradation or translation-repression via binding to mRNA 3′-UTR region in a sequence-specific manner (Valencia-Sanchez *et al.*, 2006).

Genome-wide miRNA target gene predictions suggest that post-transcriptional regulation by miRNAs is prevalent in metazoans and thousands of genes are believed to be regulated by miRNAs (Rajewsky, 2006). Therefore, integrating miRNAs into existing functional genomics data is an important step to understand the panorama of gene regulatory networks (Malphettes and Fussenegger, 2006; Rajewsky, 2006). Compared with the intensive studies that have been carried out on prediction and validation of miRNA target gene regulations, relatively little is known about the regulation of these crucial regulator themselves. Recently, several pilot experimental studies set out to uncover the transcriptional regulation of individual miRNAs. Several miRNAs are found to be controlled by specific transcription factors (TFs) that contribute to miRNA tissue- or stage-specific expression patterns (Chang *et al.*, 2004; Fazi *et al.*, 2005; Fukao *et al.*, 2007; O'Donnell *et al.*, 2005; Zhao *et al.*, 2005). For instance, *Drosophila miR-1* is reported to be controlled by the TFs like *Twist*, *Snail*, *Mef2* and *Dorsal*, which restrict its expression in mesoderm and muscle (Biemar *et al.*, 2005; Kwon *et al.*, 2005; Sokol and Ambros, 2005). However, for most of the miRNAs, the transcriptional regulatory mechanism is still unknown. Thus, computational methods are valuable and complementary to laboratory experiments to identify and characterize miRNA *cis*-regulatory elements. Up to now, only few computational studies of miRNA *cis*-acting regulatory regions have been reported in Plants (Megraw *et al.*, 2006; Wang *et al.*, 2006), worm (Ohler *et al.*, 2004) and human (Jegga *et al.*, 2007; Wu and Xie, 2006; Zhou *et al.*, 2007).

Phylogenetic footprinting (Tagle *et al.*, 1988) is a comparative genomics approach to identify *cis*-regulatory elements that are conserved in homologous sequences across multiple species (GuhaThakurta, 2006). Numerous such methods have been reported for *de novo* motif discovery. Typically, these methods can be grouped into two classes: one is alignment-based and the other is motif-based (Fang and Blanchette, 2006). The alignment-based methods start with a multiple alignment and then

*To whom correspondence should be addressed.

scan to identify conserved regions. Typically, the phylogenetically conserved elements can be detected by multiple sequence alignments in relatively closely related species. However, highly diverged sequences are difficult to align. Thus, some short conserved TFBSs that are embedded in poorly conserved regions are hard to detect between distantly related species. To overcome this shortcoming, motif-based approaches like Footprinter (Blanchette and Tompa, 2003) are developed. Such methods can detect short conserved elements but with the cost of higher false positive rate (Prakash and Tompa, 2005) and can be only applied on relatively short DNA sequences (<1 kb).

Up to now, no systematic analysis of the *cis*-regulatory elements that control miRNA expression in *Drosophila* has been reported to our knowledge. In this work, we used a two-step approach that takes advantage of both alignment-based and motif-based methods to perform phylogenetic analysis of the flanking sequences of intergenic miRNAs across 12 fly species. We first start with the pairwise alignments of the miRNA flanking sequences of the 12 *Drosophila* species, using *D.melanogaster* as the reference. Then, *D.melanogaster* sequences are scanned by a sliding window and the orthologous sequences that are aligned to *D.melanogaster* sequences of this window are analyzed with Footprinter to find conserved motifs while allowing motif duplication, deletions and rearrangement within this window. Using this approach, we analyzed the upstream 10 kb to downstream 5 kb flanking region of each known intergenic miRNA, and identified a number of CCEs across the fly species with a sensitivity of 81.8% and a false positive rate of 5.6%. These CCEs are found to be significantly enriched in binding sites of TFs that regulate development. Further analysis revealed motif pairs with significant distance constraints and overrepresented CRMs containing multiple conserved TFBSs, suggesting combinatorial miRNA gene regulation. Additionally, we identified a number of novel significantly enriched and conserved motifs in the regulatory regions of these intergenic miRNAs. Many of these motifs are also found to preferentially occur near the transcription start site (TSS) of the protein-coding genes. Finally, we tried to integrate our predictions with gene transcriptional control and miRNA target regulations, and searched for putative regulatory feedback loops of interactions between miRNAs and transcription factors.

2 MATERIALS AND METHODS

2.1 Data resource

2.1.1 MiRNAs miRNA sequences were downloaded from miRBase release 9.1 (http://microrna.sanger.ac.uk/sequences/). There are total 78 known pre-miRNAs in *D.melanogaster*.

2.1.2 Genomic sequences and Gene annotation We downloaded the FlyBaseGene annotation (updated 28 July 2006) and the genomic sequences of the 12 *Drosophila* species (Drosophila_12_Genomes_Consortium, 2007) from UCSC Genome Browser (http://genome.ucsc.edu/): *D.melanogaster (dm2)*, *D.simulans (droSim1)*, *D.yakuba (droYak2)*, *D.ananassae (droAna3)*, *D.pseudoobscura (dp4)*, *D.virilis (droVir3)*, *D.mojavensis (droMoj3)*, *D.sechellia (droSec1)*, *D. erecta (droEre2)*, *D.persimilis (droPer1)*, *D.willistoni (droWil1)* and *D.grimshawi (droGri2)*. Supplementary Figure S1 shows the phylogenetic tree of the 12 fly species.

2.1.3 Known TFBSs regulating miRNAs Eleven TFBSs that are reported to be conserved across fly species were collected from the literature (Kwon *et al.*, 2005; Sokol and Ambros, 2005). All of these sites are around *mir-1*. Nine of them are putative binding sites for *Twist* or *Snail*, one is for *Mef2* and the other is for *SRF* (see Supplementary Table S1).

2.2 Methods

2.2.1 Detect orthologous miRNAs in 12 fly species All the 78 known pre-miRNAs of *D.melanogaster* were used as queries to BLAST (NCBI blast version 2.2.6) against the genomic sequences of the other 11 species with the default settings and *E*-value cutoff = 0.1. Then, the BLAST hits were scored by miRAlign (Wang *et al.*, 2005), which is specially designed for miRNA homology searches. Compared with pure sequence alignment-based homology search methods, miRAlign further evaluates the structural conservation between the pre-miRNAs, and can identify distant homologs. The default parameters of miRAlign were used (MFE cutoff = −20 kcal/mole, minimum mature sequence identity = 70%), and the hits with similarity score ≥35 were predicted as the homologous miRNAs. Based on homologous information, we further assigned the orthologous pairs according to the following criteria: If a query pre-miRNA hits multiple homologs, only the one with the highest pre-miRNA sequence identity was taken as its putative ortholog. If the same locus shows homology to multiple query pre-miRNAs, it was assigned as the ortholog to the one with the highest sequence identity. Several ambiguous orthologous pairs were checked and adjusted manually.

2.2.2 Two-step phylogenetic footprinting method In this work, we used a strategy based on both pairwise alignments and motif-detection methods. Figure 1 shows the schematic of our approach.

(a) *Rough localization of the orthologous sequences by pairwise alignments*. We first performed BLASTZ (Schwartz *et al.*, 2003) pairwise alignments between the counterpart of the miRNA flanking sequences between *D.melanogaster* and the other 11 *Drosophila* species using the same parameters that was used by UCSC genome browser for the whole genome pairwise alignments between these species. After this step, the orthologous regions in other species were roughly aligned to the reference sequences (*D.melanogaster* sequence).

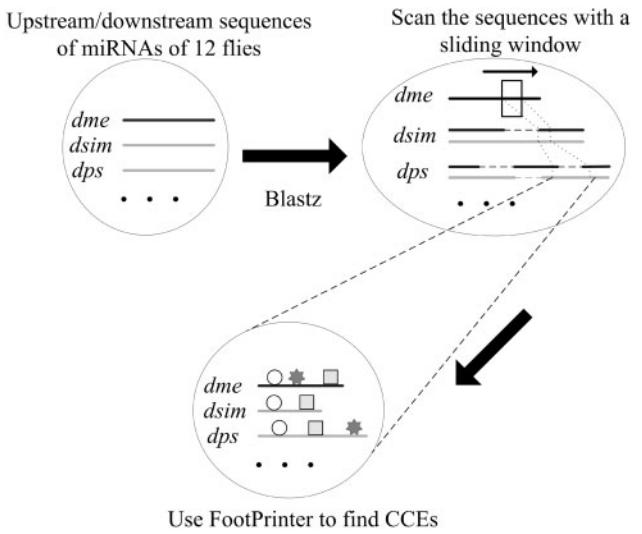

Fig. 1. Schematic representation of the two-step procedure for phylogenetic footprinting.

(b) *Detecting CCEs with Footprinter*. The reference genome was scanned by a sliding window. The reference sequence in the window and its counterpart in other species (according to the pairwise alignments) were searched using Footprinter to find CCEs. Footprinter uses a string-based motif representation to search a phylogenetic tree for motifs that show a minimal number of mismatches. This tool is very flexible so that a user can define the number of mutations that are allowed between the conserved motifs and motif losses can be handled (Blanchette and Tompa, 2002). The Footprinter parameters were set to: number of mutations allowed: 2; Maximum number of mutations per branch: 1; Motif loss cost: 1 and the low-complexity motifs were filtered with setting the parameters -filter_low_complexity1 to 0.8 and -filter_low_complexity2 to 0.99. The essential parameter of motif size will be discussed later. By default, the size of the sliding window was set to 200 nt with a step of 100 nt. Motif losses and rearrangements within the window were allowed, but motif inversions were not considered. Increasing the window size did not contribute to higher sensitivity on miRNA training set, but keeping it small can lower the amount of false positive predictions (data not shown).

One hundred random data sets with identical length of the miRNA flanking sequences were generated according to the HKY85 model (Hasegawa *et al*., 1985) for the control of false positive rate. The parameters of the HKY85 evolutionary model were estimated by PAML (Yang, 1997) using the multiple alignments of the corresponding region of the 12 fly species extracted from the UCSC genome browser.

2.2.3 Comparing to the TRANSFAC motifs To compare the known TFBSs with the CCEs identified by our approach, we scanned the CCEs with the position weight matrices (PWMs) provided by TRANSFAC 10.3 motif database (Wingender *et al*., 2001). This scanning was performed using Storm (Schones *et al*., 2007), with P-value < 0.0003 as the cutoff. A TFBS was considered to be conserved only if at least 2/3 of its site overlaps with CCEs. To estimate the number of possible TFBSs that could match purely by chance, scrambled matrices were generated with the same base composition and the same information contents as those of the true TFBS matrices by shuffling the columns of the real PWMs.

To compare 7mer motifs with the TRANSFAC motifs, we simply matched the 7mers with the consensus sequence of each known TFBSs. We eliminated consensus sequences that match too many 7mers by masking all the possible 7 nt long substring of the consensus patterns that match more than ten 7mers.

2.2.4 Identification of interacting motif pairs We used the methods introduced by Yu *et al*. (2006a) to identify the motif pairs that have significant distance constraint. The distance constraint between two motifs in the regulatory region of miRNAs was calculated by comparing the observed distance distribution between the conserved TFBSs with the background distribution using the Kolmogorove–Smirnov (KS) test. The background distribution is considered to be from motif pairs that do not interact with each other. Given a motif pair distance d, the background probability of observing d is calculated as:

$$f_d = \sum_{n=1}^{N} \frac{1}{N} \times \frac{L_n - w_f - w_s - d + 1}{\sum_{i=1}^{win} L_n - w_f - w_s - i + 1}$$

where N is the total number of input sequences, w_f and w_s are the width of the two motifs, respectively, L_n is the length of the n-th input sequence and *win* denotes the maximum distance between a motif pair. As for most of the interacting TF pairs, the distances between their binding sites are relatively short [typically ≤200 nt (Yu *et al*., 2006a, b), we arbitrarily set $win = 1000$ bp, namely only the motif pairs ≤1000 bp apart were counted.

2.2.5 Identification of cis-regulatory modules (CRMs) We used the Cumulative Conserved TFBS Score (CCTS) to identify the DNA sequences that are likely to be CRMs by detecting tight clusters of locally overrepresented conserved TFBSs. The CCTS is defined as:

$$CCTS_{\{i,j\}} = \sum_{m \in S} K_{m\_c}$$

where i and j are the start and end positions of this sequence segment, K_{m_c} is the counts of the conserved instances of motif m in this region and S denotes the motif set that contains motifs having at least two conserved instances in this region.

2.2.6 Identification of novel regulatory motifs To find putative novel regulatory motifs, we searched for the over-represented conserved 7mers in the miRNAs CCEs using the method introduced by Wu and Xie (2006) and Xie *et al*. (2005). We defined a 7mer instance to be conserved if it is located within a CCE. The enrichment of the conserved instances of each 7mer was measured using a Z-score defined as:

$$Z_i = (K_i - N_i p_0)/\sqrt{N_i p_0 (1 - p_0)}$$

where K_i and N_i are the conserved and total instances of the i-th 7mers, respectively, and p_0 is the background conservation rate of 7mers. This score measures the relative enrichment of a 7mer in the CCEs compared to the background. To achieve a significant conservation score, a 7mer must be highly conserved and overrepresented.

2.2.7 Comparison with protein-coding gene promoter sequences A set of ~6740 promoters sequences covering [−1500, +500] with respect to the TSS of protein-coding genes according to the FlyBaseGene annotation were selected to compare the microRNA regulatory sequences. If a gene has multiple TSSs, we only kept the most distal one from the CDS. In addition, the TSSs that are ≤100 bp apart from the start codon were not included. The orthologous promoter sequences of the other *Drosophila* species were extracted from the whole genome pairwise alignments generated using Mercator (Dewey, 2006) and MAVID (Bray and Pachter, 2004) (http://www.biostat.wisc.edu/~cdewey/fly_CAF1/).

2.2.8 Predictions of miRNA targets We simply searched for target protein-coding genes by identifying conserved 7mers (conserved in at least 10 different *Drosophila* species) in *D.melanogaster* 3′-UTR sequences, which are complementary to the 5′-seeds (1–7 nt or 2–8 nt) of miRNAs. To evaluate the false positives of the predictions, the seed regions (1–8 nt) of miRNA are randomly shuffled. Then the shuffled miRNAs were used to search for conserved complementary sites. We repeated the randomization for 10 times.

3 RESULTS AND DISCUSSION

3.1 Finding phylogenetically conserved *cis*-elements (CCEs) around the intergenic miRNAs

We first performed a homology search of all the 78 pre-miRNAs of *D.melanogaster* in other 11 fly species and assigned their orthologs. Seventy-one of the pre-miRNAs were detected to be conserved across all the fly species (see Supplementary Table S2). As previous works suggest that the miRNAs in the same cluster are likely to be transcribed as a polycistronic transcript, we grouped the miRNAs into clusters if (i) they are in the same intergenic region, (ii) on the same strand and (iii) the distance of adjacent pre-miRNA is ≤2000 bp. Then, the intergenic miRNAs were extracted according to the *D.melanogaster* gene annotation of FlyBaseGene. Finally, we detected 35 intergenic miRNA transcription unit candidates

that are conserved across the *Drosophila* species, consisting of 45 pre-miRNAs (see Supplementary Table S3).

Because few primary transcripts of fly miRNAs have been characterized and fly enhancers can act over long distance, we extracted a relative large region (upstream 10 kb and downstream 5 kb according to *D.melanogaster*'s genomic sequences) surrounding each miRNA. When the pre-miRNAs and its upstream/downstream protein-coding genes are unidirectional, the extracted flanking sequences of the miRNA were shortened to guarantee that it is not overlapping with any adjacent genes. The orthologous regions in other species were roughly mapped to the *D.melanogaster* genome by pairwise alignments.

Since lineage-specific motif site losses are prevalent in *Drosophila* (Moses *et al*., 2006) and some fly genomic sequences are incomplete, certain motif site losses must be allowed. To choose an appropriate parameter setting for which the sensitivity and specificity can be balanced, we performed a systematic comparison of different motif sizes and number of motif losses allowed for the Footprinter search (Table 1). As most functional TFBSs are short (~8 bp), we tested the performance of our strategy from motif size ranging from 7 to 10 nt. When we allowed three or more motif losses, to guarantee the motifs are derived from the common ancestor of these 12 fly species we required the motif to be conserved in at least one of the species of *D.virilis*, *D.mojavensis* or *D.grimshawi* (see Supplementary Fig. S1 for the phylogeny of these species). Finally, we chose motif size 8 and allowing at most 2 motif losses as our parameters for further analysis. As shown in Table 1, this choice can achieve a sensitivity of 81.8% (9/11 known conserved TFBSs correctly identified) and a false positive rate of 5.60% [(CCE length expected by chance)/(CCE length)]. The identified CCEs are available in the Supplementary Material.

3.2 Matching known TFBSs

3.2.1 Enriched known TFBSs in CCEs We compared the CCEs with the known insect TFBSs in the TRANSFAC 10.3 database (Wingender *et al*., 2001) to find possible known motifs located in the CCEs. A 100 scrambled matrices were constructed for each real PWM to scan the CCEs using the same procedure for the control. As expected, many of the CCEs match the known motifs. Using the real PWMs we detected 2555 putative TFBSs (see Supplementary Material) overlapping with CCEs which is 1.33 ± 0.06 fold higher than expected by chance (1919.4 ± 97.3 on average) (P-value < 0.01). For the comparison, the non-CE regions have only 1.04 (± 0.02) fold enrichment in TFBSs of known PWMs compared to the shuffled PWMs (Fig. 2A). Seven TFs (*Adf-1*, *Abd-B*, *Sd*, *Prd*, *Ubx*, *Zen* and *En*) have more than 2-fold putative binding sites than expected by chance (all with P-value < 0.05), which suggests many of these TFs may contribute to miRNA

Table 1. Performance with different Footprinter parameter settings

Motif losses[a]	Motif size (nt)[b]	Sensitivity[c] (%)	Detected CCE length around miRNAs (nt)[d]	Detected CCE length on random data sets (nt)[e]	FPR[f] (%)
0_losses	7	90.9	53 154	8225 ± 250	15.47
	8	54.5	33 092	2115 ± 91	6.39
	9	27.3	22 993	684 ± 73	2.97
	10	27.3	18 669	245 ± 61	1.31
2_losses	7	90.9	94 397	$11\,519 \pm 213$	12.20
	8	81.8	62 029	3476 ± 145	5.60
	9	63.6	45 288	1261 ± 74	2.79
	10	54.5	37 188	515 ± 73	1.39
3_losses	7	90.9	96 976	$11\,802 \pm 195$	12.17
	8	81.8	64 685	3645 ± 144	5.64
	9	63.6	47 826	1342 ± 90	2.81
	10	54.5	39 488	557 ± 86	1.41

[a]Motif losses in at most 0, 2 or 3 species, respectively.
[b]Motif size in nucleotide.
[c]Proportion known conserved TFBSs correctly identified.
[d]Total CCE length around the miRNAs.
[e]Average length of CCEs on 100 randomized data sets.
[f]False positive rate. FPR = (CCE length expected by chance)/(CCE length).

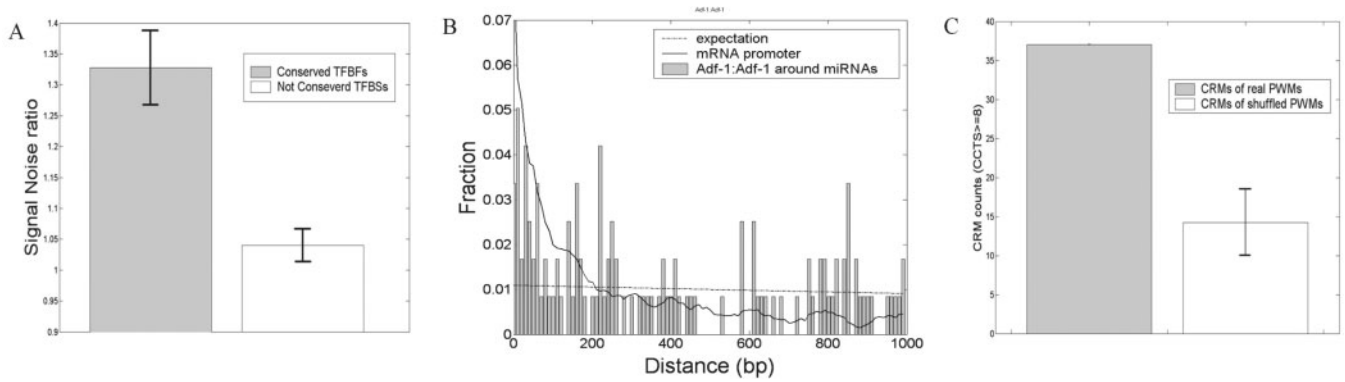

Fig. 2. Matching the known TFBSs. (**A**) CCEs are significantly enriched in known TFBSs. (**B**) Distribution of the distances between conserved Adf-1: Adf-1 binding sites around miRNAs (plotted at 10 bp intervals). The solid line indicates the distribution in the protein-coding gene promoters, and the dashed line shows the expectation by chance. (**C**) MiRNA flanking sequences are enriched in the putative CRMs. Using a CCTS cutoff of 8, we identified 36 putative CRMs. As a control, we scored these sequences with shuffled PWMs for 100 times, and only detected 14.6 (± 4.4) fake CRMs on averages.

regulation. All these TFs are related to development, and five of them belong to the homeodomain family.

3.2.2 Identification of interacting motif pairs Since pairwise TF–TF interaction is the most characteristic feature in *cis*-regulatory regions, we used the method introduced by Yu *et al.* (2006) to find the conserved TFBSs pairs with distance constraints. The argument is that if two TFs interact with each other, the distance between their binding sites is unlikely to follow a random distribution. Three motif pairs were found to reveal significant distance constraints (P-value < 0.05 after Bonferroni correction for multiple testing, KS-test). One of the significant interactions is the *Adf-1:Adf-1* self-interaction (Fig. 2B). *Adf-1* is an essential sequence-specific TF that regulates a diverse group of genes in *Drosophila*. Previous reports suggest that this TF has a protein interaction domain and may bind DNA as a dimer (Cutler *et al.*, 1998). The other two putative interacting pairs are composed by Adf-1 with transcription factor *E74A* and *Mad*, respectively (see Supplementary Fig. S2). Together with the significant overrepresentation of Adf-1 TFBSs in CCEs, this result suggests that Adf-1 may be an important transcription regulator of the *Drosophila* intergenic miRNA genes. For the comparison, we also computed the distance distribution of conserved instances of these motif pairs in the promoters of protein-coding genes. All these three motif pairs are found to have significant distance constraints in the protein-coding gene promoters (Fig. 2B and Supplementary Fig. S2). This result suggests some combinatorial controls are likely to be shared by miRNA and protein-coding genes.

3.2.3 Detecting cis-regulatory modules In eukaryotes, functional TFBSs are often found to be clustered together into CRMs (enhancers). Drosophila enhancers are typically 500–1000 bp in length and can locate far from the TSS of the regulated genes. Several previous works used the tight TFBS clustering property of the early acting transcription factors (e.g. *Bicoid*, *Hunchback*, *Krüppel*, *Knirps* and *Caudal*) to identify the putative enhancers that may be active in early Drosophila embryo (Berman *et al.*, 2002, 2004). And another recent work suggests that using the local overrepresentation property of TFBS motifs may greatly contribute to correct identification of CRMs (Pierstorff *et al.*, 2006). Here we used a CCTS score method to find potential CRMs that may regulate miRNA expression by considering local overrepresentation and conservation of all the known TRANSFAC insects TFBSs (see Methods section). To estimate the number of CRMs that could be discovered by chance, we scrambled the PWM matrices and searched the CCEs for 100 times as control. Using the criteria of at least eight local overrepresented conserved TFBSs within 1000 bp, we detected 36 potential CRMs that is about 2.47-fold higher than 14.6($\pm$4.4) identified with shuffled PWMs (P-value < 0.01) (Fig. 2C). And the previously reported proximal CRM of *mir-1* (Sokol and Ambros, 2005) are also discovered by our predictions. Supplementary Table S4 lists the predicted putative CRMs, and Table S5 shows the number of detected CRMs for the different parameter settings.

3.3 Identifying novel motifs

In the 3.2 section, we mainly focus on the conserved instances of the known motifs in CCEs. However, the majority of CCEs have no match in TRANSFAC. Thus, we sought to find novel motifs by searching for overrepresented 7mers in miRNA CCEs. For each possible 7mers, a conservation score introduced by Xie (Xie *et al.*, 2005) was calculated to measure the relative enrichment of that 7mer in CCEs compared to the background. Complementary 7mers were combined and the ones with low sequence complexity (the most common nucleotide accounts for $\geq$6 nt of the motif; di- and tri-nucleotide repeats) were not considered. About 6729 7mers were found to have at least one conserved instance around intergenic miRNAs and 119 of them are significantly enriched in miRNA CCEs with a P-value less than 10^{-6}. We reasoned that if this score had successfully identified the functional motifs that do regulate the miRNAs, then the high-scoring 7mers should have more matches to the known TFBSs than expected by chance. All these 6729 7mers were compared with the consensus sequences of the insects known motifs in the TRANSFAC database. As show in Figure 3A, highest scoring motifs significantly match more known motifs than the others and 32 of the top 100 7mers match the TRANSFAC consensus that is 3-fold higher than the average (P-value $< 10^{-7}$). The top 100 7mers matched the TFBSs of *Abd-B*, *Antp*, *byn*, *Cf1a*, *dri*, *Ftz*, *Ubx* and *Zen* for more than twice. Interestingly, all these TFs are reported to regulate Drosophila development and seven of them are

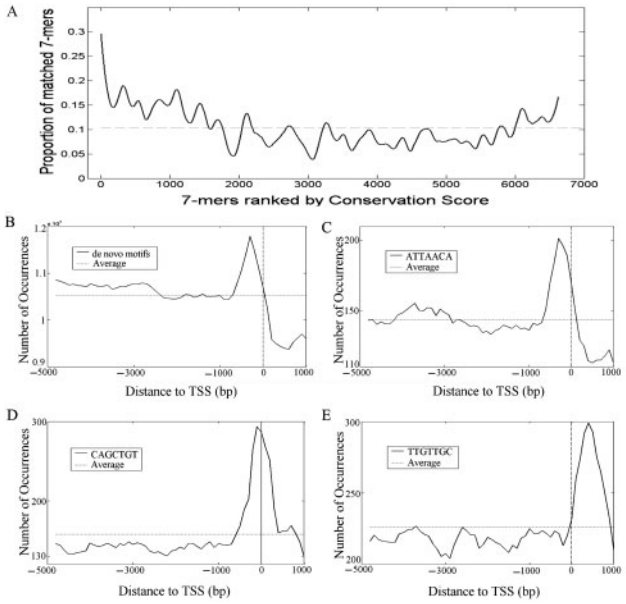

Fig. 3. Novel motifs around the intergenic miRNAs. (**A**) High scoring 7mers are significantly enriched in known TRANSFAC motifs. This figure was drawn using windows of size 200. The dashed line indicates the average portion of 7mers matching the TRANSFAC motifs. (**B**) The occurrences of the 7mers relative to TSS of protein-coding genes for all the 68 novel top 100, 7mer motifs. The motif ATTAACA (**C**), CAGCTGT (**D**) and TTGTTGC (**E**) are preferentially located at the upstream, surrounding and downstream of TSS, respectively.

homeodomain TFs. Together with the observations in Section 3.2.1, we noticed that most of the TFs predicted to be preferentially regulating the miRNAs appear to be those TFs that play essential role in drosophila development and are enriched by the members of the homeodomain family. This result coincides with the notion that most of the known *Drosophila* miRNAs are expressed in early embryo with important developmental roles. In spite of these, we were aware that some of the lowest scoring 7mers also matched known motifs (Fig. 3A). A similar phenomenon was observed in the protein-coding gene promoters too (data not show). We compared the GC content of the top 100 with the bottom 100 7mers, no significant difference was observed. One possible explanation is that the low-scoring functional elements may involve quickly, and many of these sites are linage specific.

Next, we checked whether the other 68 novel motifs of the top 100 7mers are also shared by protein-coding genes. About 37% and 94% of these motifs are also ranked in the top 100 and top 1000, respectively of the motifs identified from the protein-coding gene promoters using the same procedure. When mapping these 7mers to the promoter region of protein-coding genes, we found that many of these novel motifs are preferentially located near the TSS of protein-coding genes (Fig. 3B–E). These observations should argue for the validity and importance of these motifs and suggest that many of these miRNA 7mers may also play a role in the transcription of protein-coding genes. This observation consists with recent reports in human data (Lee *et al*., 2007). Supplementary Table S6 lists the top 100 7mers.

3.4 Constructing potential regulatory feed-back loops

Integrating miRNAs into the existing gene regulation networks is an important step towards understanding gene regulation at the systems level. Those, that involve feed-back loops of interactions among miRNAs and transcription factors, are specially interesting (Chen and Rajewsky, 2007). Several pilot experimental works demonstrate the existence and importance of such networks. One example is the reciprocal negative feedback loop between *miR-7* and transcription factor *Yan* in *Drosophila* (Li and Carthew, 2005), which ensures mutually exclusive expression with *miR-7* in photoreceptor cells and *Yan* in progenitor cells and contributes to photoreceptor differentiation. Other examples include the *miR-273*/*lsy-6*/*die-1*/*cog-1* double negative feedback loop that programs neuronal left/right asymmetry in *Caenorhabditis elegans* (Chang *et al*., 2004; Johnston *et al*., 2005; Poole and Hobert, 2006), the interaction between *NFI-A* and *miR-223* plays a crucial role in granulopoiesis (Fazi *et al*., 2005) and the regulatory network composed by *c-MYC*, *E2F* and *miR-17* cluster that may regulate cellular proliferation and apoptosis in human (O'Donnell *et al*., 2005; Sylvestre *et al*., 2007). Thus, based on the conserved putative known TFBSs identified above, we further searched for potential regulatory feedback loops of the form:

$$TF_{start} \rightarrow \text{miRNAs} \rightarrow (\text{mRNA} \rightarrow TF_{end})$$

where $TF_{start} = TF_{end}$. In addition, we tried to estimate the number of false positive predictions by detecting such feedback loops with the shuffled PWMs and miRNAs. Our result suggests that out of the 18 predicted feedback loops (see Supplementary Table S7) 5.6($\pm$2.1) may be false-positives. One of these putative feedback loops is composed by *miR-1* and *Su(H)*. *Su(H)* is an important component of Notch signaling pathway and promotes the differentiation of pericardial cells in *Drosophila*. While *miR-1*, which targets the Notch signaling pathway and contributes to muscle development, is found to be expressed in myocardial cells but not in pericardial cells. A putative reciprocal negative feedback loop between *miR-1* and *Su(H)* is speculated to reinforce proper differentiation of cardiac cells (Sokol and Ambros, 2005).

4 CONCLUSION

In this work, we used a two-step phylogenetic footprinting strategy to analyze the flanking sequences of intergenic miRNAs in *Drosophila*. This approach takes the advantages of high sensitivity of motif-detection methods and reduced false-positive rate by rough localization of orthologous sequences through pairwise alignments. In principle, our method is similar to Footprinter3.0 (Fang and Blanchette, 2006) that also starts with sequence alignments and ends with Footprinter search. But Footprinter3.0 only provides web services and can only be applied on relatively short sequences like core promoters.

Using this two-step method, we identified a number of putative miRNA gene *cis*-regulatory elements, which are significantly enriched with the binding sites of the TFs that regulate development. Based on these CCEs, we further identified motif pairs that have significant distance constraints, CRMs consisting of multiple TFBSs and a number of novel motifs, many of which preferentially occur near the TSS of the protein-coding genes. Additionally, we tried to integrate our predictions with known functional genomics data and searched for putative feedback loops of interactions between miRNAs and transcription factors. These results have extended the existing knowledge on transcriptional regulation of *Drosophila* miRNAs, and provide a foundation for further studying of miRNAs' role in *Drosophila* gene regulatory networks. While, we are aware that both predictions of TFBSs and microRNA targets have certain level of positive rates and our identified feedback loops have not yet been validated experimentally. (As *miR-7* is hosted in a protein-coding gene, it was not included in our intergenic miRNA set.) During the revision of this article, (Tsang *et al*., 2007) reported a computational analysis of the potential feedback and feedforward loops between intronic miRNAs and their target genes in mammals. They found such regulatory loops are prevalent in human and mouse, and may have a role in enhancement of the robustness of gene regulation. Further computational and experimental validation and investigation of miRNA mediated regulatory loops are necessary in the future in order to fully understand miRNAs' function and the gene regulation at a systems level.

ACKNOWLEDGEMENTS

We thank Mr Tao Peng and Mr Yu Liu for helpful discussion. This project is supported in part by NSFC (grants 60234020, 60572086), 863 Hi-tech research and development program of P. R. China (No.2006AA02Z311) and the China National

Technology Platform. M.Q.Z. is also partly supported by the Chang Jiang Scholarship Program and by NIH HG06916.

Conflict of Interest: none declared.

REFERENCES

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Berman,B.P. *et al*. (2002) Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the Drosophila genome. *Proc. Natl Acad. Sci. USA*, **99**, 757–762.

Berman,B.P. *et al*. (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. *Genome Biol.*, **5**, R61.

Biemar,F. *et al*. (2005) Spatial regulation of microRNA gene expression in the Drosophila embryo. *Proc. Natl Acad. Sci. USA*, **102**, 15907–15911.

Blanchette,M. and Tompa,M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**, 739–748.

Blanchette,M. and Tompa,M. (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.*, **31**, 3840–3842.

Borchert,G.M. *et al*. (2006) RNA polymerase III transcribes human microRNAs. *Nat. Struct. Mol. Biol.*, **13**, 1097–1101.

Bray,N. and Pachter,L. (2004) MAVID: constrained ancestral alignment of multiple sequences. *Genome Res.*, **14**, 693–699.

Cai,X. *et al*. (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, **10**, 1957–1966.

Chang,S. *et al*. (2004) MicroRNAs act sequentially and asymmetrically to control chemosensory laterality in the nematode. *Nature*, **430**, 785–789.

Chen,K. and Rajewsky,N. (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.*, **8**, 93–103.

Cutler,G. *et al*. (1998) Adf-1 is a nonmodular transcription factor that contains a TAF-binding Myb-like motif. *Mol. Cell Biol.*, **18**, 2252–2261.

Dewey,C.N. (2006) Whole-genome alignments and polytopes for comparative genomics. *Ph.D. Thesis*. University of California, Berkeley.

Drosophila_12_Genomes_Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.

Fang,F. and Blanchette,M. (2006) FootPrinter3: phylogenetic footprinting in partially alignable sequences. *Nucleic Acids Res.*, **34**, W617–W620.

Fazi,F. *et al*. (2005) A minicircuitry comprised of microRNA-223 and transcription factors NFI-A and C/EBPalpha regulates human granulopoiesis. *Cell*, **123**, 819–831.

Fukao,T. *et al*. (2007) An evolutionarily conserved mechanism for microRNA-223 expression revealed by microRNA gene profiling. *Cell*, **129**, 617–631.

GuhaThakurta,D. (2006) Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res.*, **34**, 3585–3598.

Hasegawa,M. *et al*. (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, **22**, 160–174.

Jegga,A.G. *et al*. (2007) GenomeTrafac: a whole genome resource for the detection of transcription factor binding-site clusters associated with conventional and microRNA encoding genes conserved between mouse and human gene orthologs. *Nucleic Acids Res.*, **35**, D116–D121.

Johnston,R.J. Jr *et al*. (2005) MicroRNAs acting in a double-negative feedback loop to control a neuronal cell fate decision. *Proc. Natl Acad. Sci. USA*, **102**, 12449–12454.

Kwon,C. *et al*. (2005) MicroRNA1 influences cardiac differentiation in Drosophila and regulates Notch signaling. *Proc. Natl Acad. Sci. USA*, **102**, 18986–18991.

Lee,J. *et al*. (2007) Regulatory circuit of human microRNA biogenesis. *PLoS Comput. Biol.*, **3**, e67.

Lee,Y. *et al*. (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, **23**, 4051–4060.

Li,X. and Carthew,R.W. (2005) A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the Drosophila eye. *Cell*, **123**, 1267–1277.

Malphettes,L. and Fussenegger,M. (2006) Impact of RNA interference on gene networks. *Metab. Eng.*, **8**, 672–683.

Megraw,M. *et al*. (2006) MicroRNA promoter element discovery in Arabidopsis. *RNA*, **12**, 1612–1619.

Moses,A.M. *et al*. (2006) Large-scale turnover of functional transcription factor binding sites in Drosophila. *PLoS Comput. Biol.*, **2**, e130.

O'Donnell,K.A. *et al*. (2005) c-Myc-regulated microRNAs modulate E2F1 expression. *Nature*, **435**, 839–843.

Ohler,U. *et al*. (2004) Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA*, **10**, 1309–1322.

Pierstorff,N. *et al*. (2006) Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics*, **22**, 2858–2864.

Poole,R.J. and Hobert,O. (2006) Early embryonic programming of neuronal left/right asymmetry in C. elegans. *Curr. Biol.*, **16**, 2279–2292.

Prakash,A. and Tompa,M. (2005) Discovery of regulatory elements in vertebrates through comparative genomics. *Nat. Biotechnol.*, **23**, 1249–1256.

Rajewsky,N. (2006) microRNA target predictions in animals. *Nat. Genet.*, **38** (Suppl.), S8–S13.

Schones,D.E. *et al*. (2007) Statistical significance of cis-regulatory modules. *BMC Bioinformatics*, **8**, 19.

Schwartz,S. *et al*. (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.

Sokol,N.S. and Ambros,V. (2005) Mesodermally expressed Drosophila microRNA-1 is regulated by Twist and is required in muscles during larval growth. *Genes Dev.*, **19**, 2343–2354.

Sylvestre,Y. *et al*. (2007) An E2F/miR-20a autoregulatory feedback loop. *J. Biol. Chem.*, **282**, 2135–2143.

Tagle,D.A. *et al*. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.*, **203**, 439–455.

Tsang,J. *et al*. (2007) MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. *Mol. Cell*, **26**, 753–767.

Valencia-Sanchez,M.A. *et al*. (2006) Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev.*, **20**, 515–524.

Wang,X. *et al*. (2005) MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, **21**, 3610–3614.

Wang,Y. *et al*. (2006) Significant sequence similarities in promoters and precursors of Arabidopsis thaliana non-conserved microRNAs. *Bioinformatics*, **22**, 2585–2589.

Wingender,E. *et al*. (2001) The TRANSFAC system on gene expression regulation. *Nucleic Acids Res.*, **29**, 281–283.

Wu,J. and Xie,X. (2006) Comparative sequence analysis reveals an intricate network among REST, CREB and miRNA in mediating neuronal gene expression. *Genome Biol.*, **7**, R85.

Xie,X. *et al*. (2005) Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature*, **434**, 338–345.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.

Yu,X. *et al*. (2006a) Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **34**, 917–927.

Yu,X. *et al*. (2006b) Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.*, **34**, 4925–4936.

Zhao,Y. *et al*. (2005) Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature*, **436**, 214–220.

Zhou,X. *et al*. (2007) Characterization and identification of MicroRNA core promoters in four model species. *PLoS Comput. Biol.*, **3**, e37.