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

Xiaowo Wang¹, Jin Gu¹, Michael Q. Zhang¹,² and Yanda Li¹,*

¹MOE Key Laboratory of Bioinformatics and Bioinformatics Division, TJNUST/Department of Automation, Tsinghua University, Beijing 100084, China and ²Cold 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

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 metazoa 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 mRNA 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 (Megrav 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
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 (<1kb).

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.

## 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* (dsim1), *D.yakuba* (dyYak2), *D.ananassae* (dnaA3), *D.pseudoobscura* (dps), *D.virilis* (dvVir3), *D.mojavensis* (dmOj3), *D.sechellia* (dSce1), *D. erecta* (deEre2), *D.persimilis* (dPer1), *D.willistoni* (dWill) and *D.grimshawi* (dGri2). 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 miRAliign (Wang _et al._, 2005), which is specially designed for miRNA homology searches. Compared with pure sequence alignment-based homology search methods, miRAliign further evaluates the structural conservation between the pre-miRNAs, and can identify distant homologs. The default parameters of miRAliign 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*).

(b) **Upstream/downstream sequences of miRNAs of 12 flies** Scan the sequences with a sliding window. Use Footprinter to find CCEs.

Fig. 1. Schematic representation of the two-step procedure for phylogenetic footprinting.
**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 as having the maximum distance between a motif pair. As for the most of the interacting TF pairs, the distances between their binding sites are relatively short ([typically ≤200 nt](Yue 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:

$$CCTS_{i, j} = \sum_{m=1}^{S} K_{m,c}$$

where $i$ and $j$ are the start and end positions of this 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.

#### 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 as having the maximum distance between a motif pair. As for the most of the interacting TF pairs, the distances between their binding sites are relatively short ([typically ≤200 nt](Yue et al., 2006a, b), we arbitrarily set win = 1000 bp, namely only the motif pairs ≤1000 bp apart were counted.
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 expression.

**Table 1. Performance with different Footprinter parameter settings**

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

*Motif losses in at most 0, 2 or 3 species, respectively.*

*Motif size in nucleotide.*

*Proportion known conserved TFBSs correctly identified.*

*Total CCE length around the miRNAs.*

*Averaged CCE length on 100 randomized data sets.*

*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, Knüppel, 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±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 ≥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^-6). The top 100 7mers matched the TFBSs of Abd-B, Antp, bby, Cfi1a, dri, Fiz, 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 TTGTGCC (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 lineage 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 feedback 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/Isy-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; Sylvester et al., 2007). Thus, based on the conserved putative known TFBSs identified above, we further searched for potential regulatory feedback loops of the form:

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

where $\text{TF}_{\text{start}} = \text{TF}_{\text{end}}$. 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(±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 orthologues 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, CRM 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


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


