The regulatory epicenter of miRNAs

Ashwani Jha[†], Mrigaya Mehra and Ravi Shankar[†],*

Studio of Computational Biology & Bioinformatics, Biotechnology Division, CSIR-Institute of Himalayan Bioresource Technology, Council of Scientific & Industrial Research, Palampur 176 061, HP, India

*Corresponding author (Email, ravish@ihbt.res.in, ravish9@gmail.com)

[†]Equal contribution

miRNAs are small non-coding RNAs with average length of ~21 bp. miRNA formation seems to be dependent upon multiple factors besides Drosha and Dicer, in a tissue/stage-specific manner, with interplay of several specific binding factors. In the present study, we have investigated transcription factor binding sites in and around the genomic sequences of precursor miRNAs and RNA-binding protein (RBP) sites in miRNA precursor sequences, analysed and tested in comprehensive manner. Here, we report that miRNA precursor regions are positionally enriched for binding of transcription factors as well as RBPs around the 3′ end of mature miRNA region in 5′ arm. The pattern and distribution of such regulatory sites appears to be a characteristic of precursor miRNA sequences when compared with non-miRNA sequences as negative dataset and tested statistically. When compared with 1 kb upstream regions, a sudden sharp peak for binding sites arises in the enriched zone near the mature miRNA region. An expression-data-based correlation analysis was performed between such miRNAs and their corresponding transcription factors and RBPs for this region. Some specific groups of binding factors and associated miRNAs were identified. We also identified some of the over-represented transcription factors and associated miRNAs with high expression correlation values which could be useful in cancer-related studies. The highly correlated groups were found to host experimentally validated composite regulatory modules, in which Lmo2-GATA1 appeared as the predominant one. For many of RBP–miRNAs associations, co-expression similarity was also evident among the associated miRNA common to given RBPs, supporting the Regulon model, suggesting a common role and common control of these miRNAs by the associated RBPs. Based on our findings, we propose that the observed characteristic distribution of regulatory sites in precursor miRNA sequence regions could be critical in miRNA transcription, processing, stability and formation and are important for therapeutic studies. Our findings also support the recently proposed theory of self-sufficient mode of transcription by miRNAs, which states that miRNA transcription can be carried out in host-independent mode too.

1. Introduction

In the present era, non-coding elements of genome have gained a new found respectability with their emergence as king-makers of genomes and their involvement in genome dynamics, evolution and regulation. miRNAs are one of such elements that have proved themselves immensely critical in the process of regulation and control of several cellular processes (Zhang *et al.* 2006; Bushati and Cohen 2007; Stefani and Slack 2008). The first discovered case of miRNA regulation was

identified for the negative regulatory impact *lin*-4 has on *lin*-14 gene, a requirement for the normal temporal development of *Caenorhabditis elegans* (Lee *et al.* 1993). miRNAs are single-stranded RNA molecules having a length of approximately 21 to 25 nucleotides. They are generated endogenously from transcripts having hairpin structures. The major function attributed to miRNAs is to reduce the expression level of targeted genes either through cleavage of target mRNAs or translation disruption (Liu *et al.* 2008; Filipowicz *et al.* 2008). Post transcription, miRNAs are released as long primary

Supplementary materials pertaining to this article are available on the *Journal of Biosciences* Website at *http://www.ias.ac.in/jbiosci/Sep2011/pp621–638/suppl.pdf*

transcripts, pri-miRNAs, having one or more hairpin structures with long single-stranded 5′ and 3′ terminal regions. These pri-miRNA transcripts are processed by Drosha-DGCR8 microprocessor complex into hairpin-shaped pre-miRNAs, having an average length of ~70–90 bases or higher, with 3′ overhangs. The microprocessor complex has two RNase-III-binding and double-stranded RNA-binding domains (Han *et al.* 2004). Once in the cytoplasm, the pre-miRNA is taken up by another RNase III enzyme, Dicer, which cleaves pre-miRNA and releases the miRNA duplex for mRNA targeting. One of the strands of the duplex is incorporated into the Argonaute complex, which guides it to the target site, down-regulating the target gene. Therefore, the entire process of miRNA biogenesis has two major points of regulation. The first major control point is the process of precursor sequence transcription, carried out by RNA polymerase II (Lee *et al.* 2004) as well as RNA polymerase III (Borchert *et al.* 2006), while a series of control points follow after transcription in the form of steps involved in precursor RNA processing, involving Drosha and Dicer RNAse III enzymes (Siomi and Siomi 2010). Like other genes, transcription of miRNAs is influenced by transcription factors. Studies as early as the one on temporal expression of let-7 have proved this with transcription factor TRAF regulating let-7 transcription by binding to upstream genomic region of let-7 (Johnson *et al.* 2003). Transcription factors like Myc and E2F have been observed to regulate many miRNAs, controlling cell cycle through positive and negative feedback loops with target miRNAs (Aguda *et al.* 2008). Similarly, some studies have reported p53 as a master regulator for many miRNAs like mir-34a (Tarasov *et al.* 2007; Chang *et al.* 2007). The feed-forward loop of miRNA regulation by transcription factors has been identified as the most common regulatory loop (Shalgi *et al.* 2007). Based on previous studies, it has been believed that transcription of miRNA is context specific, as intergenic miRNAs have their own promoters to transcribe while intragenic/intronic miRNAs depend upon their hosts for transcription (Rodriguez *et al.* 2004). However, there have been some very recent developments in this field which suggest a need to revise our approach towards understanding miRNA biogenesis and turn the focus on precursor miRNA sequences. According to Berezikov's group, contrary to the widespread notion that intronic miRNA are transcribed by the host gene transcription system, at least one-third of intronic miRNA in *C. elegans* were observed to transcribe independently (Isik *et al.* 2010). Very recently, a series of experimental studies by Gao *et al.* (2010) have revealed some fascinating and surprising findings regarding miRNA transcription by reporting innate capability of precursor miRNA regions to carry out autonomous transcription. They found that independent precursor miRNA sequences were able to transcribe themselves without any external promoter upstream. The precursor sequences owe this to their internal

sequences and the process of transcription behaves in a stage- and tissue-specific manner (Gao *et al.* 2010). This particular study attracted our attention to the need to look into the sequences of precursor miRNAs for their potential to undergo transcription and house regulatory sockets. Very few studies have been carried out in this context as, generally, the 5′ upstream regions or host promoter system have been studied in order to understand miRNA transcription. On the other hand, recent developments in understanding precursor miRNA processing have made some major revelations, shifting the focus from Drosha and Dicer to several other RBPs. At this juncture, the case of temporal expression of let-7 miRNA deserves mentioning as it may be considered one of the critical studies on biogenesis in which recent findings have revealed the role of a precursor miRNA interacting factor, Lin28, in controlling Drosha processing. As already mentioned above, controlled expression of mature let-7 is required for normal development in *C. elegans*. It was discovered that transcription and expression of let-7 gene was being controlled by a temporal regulatory element, located upstream of the let-7 precursor region (Johnson *et al.* 2003). The authors also reported that despite expression of precursor let-7 RNA, mature let-7 was expressed only at specific stages, depending upon Lin28, a factor required for the correct timing of mature let-7 expression. However, since the mechanism of action of Lin28 was not clear for this case, it was assumed that Lin28 could bind to the upstream genomic region to regulate expression in association with TREF. Later, through a study conducted by Newman *et al.* (2008), it was found that the stem loop region of let-7 was extremely conserved and had binding sites for some RBPs. It was found that Lin-28 actually binds to the pri-miRNA sequence's stem loop region and hinders processing by Drosha–DGCR8 complex (Newman *et al.* 2008). At present, many factors like SMAD, Lin28, hnRNPs, etc., have been found responsible in controlling miRNA biogenesis through binding to the precursor miRNA instead of DNA (Siomi and Siomi 2010). In mammals about 20 hnRNP proteins exists, which display wide range of activities through binding to the RNAs. Although hnRNP A family has been found to regulate splicing, some recent studies have found hnRNP A1 associated with miRNA processing, binding to precursor miRNA stem and loop regions, interacting with Drosha for miRNA maturation process (Guil and Cáceres 2007). Later, Michlewski *et al.* (2008) conducted a RNA-affinity-purification-based study and found that precursor miRNAs might interact with several hnRNPs which could control post-trancriptional fate of miRNAs in specific and different ways. SMAD proteins have been found to bind the stem region of pre-miRNA-21 and control Drosha-mediated miRNA maturation in response to TGF-beta signalling, for control of vascular smooth muscle cell phenotype (Davis *et al.* 2010).

Pumilio proteins are a family of homologous RNA-binding proteins (RBPs) that have been found to be associated with many RNAs, performing array of different roles, being suppressor, activators as well as transporters, in a highly contextual manner (Wickens *et al.* 2002; Kaye *et al.* 2009; Braines 2005; Gerber *et al.* 2004). Lately, Pumilio proteins were discovered to participate in association with some miRNAs. Some miRNAs share binding sites with Pumilio proteins which could have different associated roles including the possible one involving binding of Pumilio to miRNA and guiding it to bind the target, very similar to the mechanism adopted by Argonaute proteins (Nolde *et al.* 2007). A few recent studies have reported that Drosha knockout in mice affected the expression of only few miRNAs while p68/p72 knockout proved lethal. It was found that p68/p72 complex interacts with the precursor sequence to control miRNA processing (Stark *et al.* 2008; Fukuda *et al.* 2007). A recent study has reported that, rather than Dicer, it is Argonaute complex that interacts with pre-miRNAs to process out mature miRNAs (Cheloufi *et al.* 2010).

All these recent findings and advancements certainly put the precursor miRNAs in the centre stage for regulatory studies. Encouraged with this and in order to address the above-mentioned possible regulatory issues associated with precursor sequences, we carried out this study to look into the precursor miRNAs in regulatory context and identify the associated regulatory elements. We expected to find some reasoning for the recently made observations regarding transcription and processing of miRNAs through this study. Very few computational works have been done in this direction, as performing computational studies requires some experimental data backup, which has been otherwise scarce till recently. Transcription-factor-based data and related tools for many species are available in satisfactory number, which could be useful in detecting experimentally validated regulatory elements, as well as detect their distribution pattern and composite regulatory modules hidden inside the precursor sequences. However, very small amount of information existed on RBPs and their response elements/binding sites till recently, limiting RBP-associated studies. The current study carefully utilized the existing experimentally validated resources on transcription factors and RBPs and harnessed available expression data to find out the possible regulatory repertoire residing inside the precursor miRNAs, which could influence miRNA biogenesis, stability, life and transportation.

2. Materials and methods

2.1 *Sequence data*

We downloaded pre-miRNA and mature miRNA sequences for all of the 1048 human miRNAs from miRBase. We also downloaded all 213 pre-miRNA and mature miRNA sequences available for *Arabidopsis thaliana*, from MiR-Base (Griffiths-Jones *et al.* 2008), version 16.0, released on September 2010. UCSC genome browser and Ensemble were used to locate the genomic positions of miRNAs and elements around them. Negative dataset for pseudo-miRNAs were collected from the dataset used by Sewer *et al.* (2005) and Yousef *et al.* (2006). These datasets have been used frequently by groups working on miRNA identification, and are considered a reliable negative datasets. Another set of sequences was formed containing 1 kb upstream sequence regions of precursor miRNA along with the associated precursor miRNA sequences, using UCSC table browser.

2.2 *TFBS analysis*

We used Transfac database (Wingender *et al.* 1996) to map transcription factors on human sequences, applying core similarity and matrix similarity thresholds of 0.95 and 0.9, respectively, along with the option to minimize false positives and false negatives. Only high-quality matrices were considered in the study. We mapped the TFBS on *Arabidopsis* sequences using PLACE database (Higo *et al.* 1999) version 26.0. We mapped the coordinates of mature miRNA and TFBS on the result files generated for the precursor sequences.

2.3 *Composite regulatory module detection*

Composite/coregulatory modules (CRMs) are transcriptional functional units composed of two regulatory elements together, promoting transcription in synergistic fashion. A set of pre-miRNA sequences and associated transcription factors were studied for detection of composite regulatory modules, using cis-RED (Robertson *et al.* 2006) as well as TransCompel (Matys *et al.* 2006). Both of these databases have their inbuilt tools to scan for regulatory modules. The cutoff *P*-value was kept as 0.05. The reported regulatory modules are archived in these databases from experimentally validated and annotated data.

2.4 *RBP binding sites analysis*

Recently a repository for RBPs has been made available with RNA Binding Protein Database (RBPDB) (Cook *et al.* 2010). RBPDB has experimentally validated RBP–RNA interaction data for four species: human, mouse, nematodes and drosophila. So far, a total of 1456 different experiments have been included in it. For human, there are total 424 reported cases. RBPDB provides a scanning tool to identify RBP binding sites in the given query sequences, applying

Position Weight Matrices (PWM) in the backdrop. At present, there are only 73 PWM available in RBPDB, derived from binding and interaction data for various RBPs. We applied a stringent relative score cutoff of 0.9 in order to minimize the false positives while mapping the RBPs on miRNA sequences. RBPDB provides standalone version package for large amount of data analysis. For our study we downloaded and installed the standalone version on an Ubuntu Linux workstation.

2.5 *Prediction of RNA secondary structure*

We built the secondary structures of pre-miRNAs using RNAfold software (Hofacker 2003) from Vienna package and mapped the coordinates for loop start and end with respect to the mature miRNA position in the sequence and structure. For a few cases, the structures were not in concordance with those reported in MiRBase. For such instances, we considered structures reported in MiRBase as we adhered to structures reported in MiRBase. The binding sites for RBPs and TFBS were mapped with respect to the mature miRNA and hairpin loop structure, while their various statistics were calculated simultaneously.

2.6 *Reference point for binding site mapping and plotting*

After the final processing, the mature miRNA is released as a double-stranded element, having one strand from the 5′ arm and the other one from 3′ arm of the precrusor sequence. The selection of suitable miRNA candidate for targeting was decided by many factors, and there are several cases in which both the strands function as targeting miRNAs. In case of human, out of 1048 miRNA sequences, ~800 sequences were from 5′ arm of the precursor sequences. For the sake of clarity, we considered the 3′ end of the miRNA region in 5′ arm as the reference point for distance mapping. Even for the instances of miRNA from 3′ arm of the precursor sequences, their 5′ ends were mapped to their corresponding 3′ pairing region, to keep a uniform reference point for all cases. Henceforth, we will use the term reference position to mean the same. Distribution plotting for binding sites was done with respect to the reference position.

2.7 *Transcription start site analysis*

In order to measure the transcription potential of precursor sequences, we performed transcription start site (TSS) propensity measurement, using two most accurate TSS recognition tools (Abeel *et al.* 2009), ARTS and Eponine, designed for RNA-Pol-II-dependent systems. ARTS applies the well-proven accuracy of the Kernel method of mapping,

using various features like TSS signal, k-mer spectrum for regulatory/promoter region, exonic spectrum, twist angles, thermodynamic, stabilization energy and first downstream region properties (Sonnenburg *et al.* 2006). Eponine uses Relevance Vector Machine and probabilistic modelling to measure TSS potential, using major properties of RNA-Pol-II-transcribed genes, i.e. GC density distribution and TATA box (Down and Hubbard 2002).

2.8 *Expression data analysis*

Through analysing plots between binding site distance from the reference position and binding site frequency, we restricted our study to the critical region mostly falling in range of −20 bp upstream to +20 bp downstream from the reference point. Usually this region covers the entire mature microRNA region, and a part of the loop region as well was found symmetrically placed around the enriched region and reference position (see the result section for details). To calculate expression-profile-based similarity and correlation, we used recently released tool, mimiRNA, at *http://mimirna.centenary.org.au/mep/formulaire.html* (Ritchie *et al.* 2009). mimiRNA has expression data for 635 human miRNAs over 188 tissues and cell types. We calculated the correlation coefficients for each miRNA and its associated TFBS and RBPs (supplementary data 1) for the preferred region. Besides this, we studied 70 pre-miRNAs for their genomic locations, and computed expression correlation coefficient for pre-miRNAs which were in some host gene.

2.9 *Statistical analysis*

Significance estimation through *P*-values measurement for TFBS and RBPs enrichment was done using multiple binomial tests using R package. R package Mann–Whitney test was used to find the significance and exclusivity of distribution of binding sites in the precursor miRNA sequences. This analysis involved statistical comparison between the distribution patterns found for TFBS and RBPs sites in human precursor miRNA sequences and ~1000 pseudo-precursor miRNA sequences used as the negative dataset by previous studies on miRNA prediction and classification (supplementary data 3). For RBPs, all pre-miRNA sequences reported bindings sites, while for TFBS, only 540 negative dataset sequences exhibited binding sites. Considering this, in order to avoid class imbalance in the case of TFBS distribution test, we considered only those sequences that had TFBS sites and took equal number of negative (non-pre-miRNA sequences) and positive instances (pre-miRNA sequences), limited by the number of negative dataset instances. Chi-square goodness-of-fit test was carried out to estimate the significance of observed enrichment of transcription

factor–miRNA associations in the category having correlation coefficient values above 0.5, considering the distribution for all six categories. Significance for distance-specific enrichment for TFBS and RBPs was estimated by applying the Poisson distribution. Distribution of all TFBS and RBP binding sites along the full-length precursor miRNA was measured for every precursor miRNA, with respect to the reference position. The λ value was calculated from the overall occurrence frequencies calculated for each position, normalized by total positions considered for frequency estimation.

3. Results

3.1 *Precursors of miRNAs contain a characteristic region, enriched for binding of transcription factors as well as RNA-binding proteins*

In a computational analysis of all presently known miRNAs for human available in MiRBase, we found that there is an enriched region for binding sites for transcription factors as well as RBPs in the precursor sequences of miRNAs. This preferred enriched region centers mainly around the reference position. Regardless of their differences in control mechanisms, transcription factors and RBPs both show a characteristic commonality by having preference for the binding sites in the above-mentioned region in the miRNA precursor sequences. This was evident for most of the miRNAs. Over 2000 TFBS were mapped across 1048 human miRNA sequences, whose distribution was recorded for full-length sequences as well as with respect to the end of the reference position in the given precursor sequences. The same was done for more than 200 experimentally reported and validated RBPs and their binding sites. A clear skew was evident for the region around the reference position (figure 1). Interestingly, most of the miRNAs were noted to be present in this region with some transcription factor binding site, which was preferentially enriched for many TFBS, positionally, when compared against the rest of the regions. The same was observed for various RBPs. Detailed information about each miRNA, their associated TFBS, relative positions, etc., are given in supplementary data 2. Distance region −20 to 0 usually overlaps with mature miRNA regions, while +1 to +20 bases distance range covers a portion of loop region too. The highest frequency of regulatory sites was observed around distance of +1 for transcription factors. For transcription factors, position −18 and +11 recorded the steepest fall with highest values for the tangent. Similar analysis was performed over more than 200 known precursor miRNA sequences of *Arabidopsis*, in order to verify if the observed phenomenon was evident for plants too. Around 202 unique transcription factors from PLACE database were mapped over the

precursor miRNA sequences of *Arabidopsis*. The plot for distribution of transcription factor binding sites against the distance from reference position was very similar to the one observed for human sequences (figure 1).

As the observed enrichment in the preferred region could be a characteristic feature of precursor miRNAs, able to discriminate them from non-miRNAs, a Mann–Whitney test was performed. A negative datatset (supplementary data 3) comprising sequences for pseudo-miRNAs, UTR regions and t-RNA-like elements was used. This negative dataset has been reported and used already many times in the past for various miRNA characterization studies (Sewer *et al.* 2005; Yousef *et al.* 2006). Using RBPDB and TRANSFAC, negative dataset sequences were scanned for regulatory sites and distribution for each factor in each sequence as well as distribution at each sequence position was recorded. The distribution was compared with the distribution obtained for pre-miRNA using the Mann–Whitney non-parametric test. Compared with precursor miRNA, the negative set had almost uniform and random distribution for all positions. Also, the abundance was comparatively lesser for the similar regions. We found that the observed binding pattern was significantly exclusive for precursor miRNA sequences with *P*-value << 0.001 (figure 2). Although the distribution pattern for RBP as well as TFBS in pre-miRNA sequences both showed highly significant difference from non-miRNA sequences, the difference was much more stark for TFBS binding patterns. In an another set of analysis, we analysed 1 kb upstream regions of miRNA precursor sequences, along with their respective precursor sequences, for distribution of binding sites. A uniform distribution of binding sites was observed for the upstream region, almost like the one observed for the negative dataset used in the above-mentioned analysis. A sudden sharp peak appeared in the precursor miRNA sequences region, which centered around the preferred region (figure 3), corroborating our earlier made observation that the precursor miRNA sequences host a characteristic enriched region for regulatory sites. In order to know if the observed pattern was biased for miRNAs with respect to their residence status, precursors were also analysed with respect to their genomic location, i.e. (1) those residing inside the intronic regions and (2) miRNAs belonging to intergenic regions. No difference in distribution pattern was observed for both of the categories, implying the global status of the observed enrichment pattern, specific to miRNAs (supplementary data 5, figure 1).

A separate analysis was conducted for regions containing mirtrons. Mirtrons are special cases of miRNA biogenesis, mainly observed in nematodes and drosophila (Ruby *et al*. 2007). However, Berezikov *et al.* (2007) reported around 14 mirtrons in human. Due to two distinct features, mirtrons become relevant for this study: (1) Unlike regular miRNAs, mirtrons are Drosha processing independent. The stem
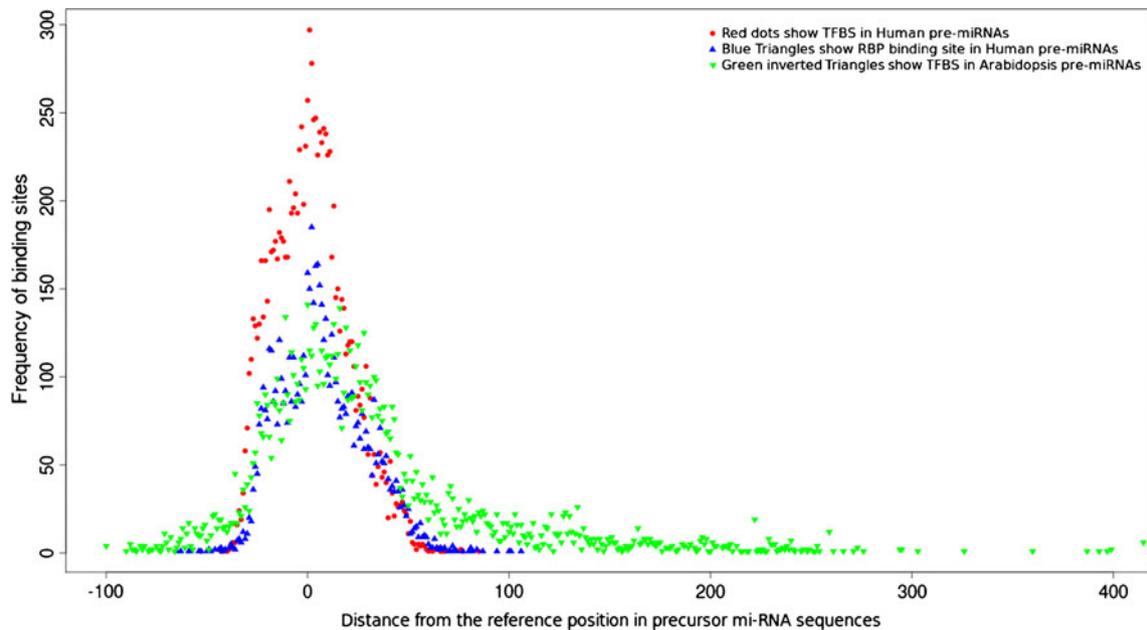
Figure 1. Distribution plot for binding factors with respect to distance from the end of mature miRNA region proximal to terminal loop in pre-miRNA sequences. All the three plots exhibit a characteristic peak in the enriched region, having very significant *P*-values for enrichment.

region of mirtronic precursors contains independent regions capable of undergoing splicing and bypasses the need of Drosha to chop off the same region. (2) Mirtrons are situated in the intronic regions of host genes and are produced as a by-product of host transcript splicing, making them dependent upon the host transcription process. Although there are only few mirtrons in human, analysing mirtrons for the binding region distribution could shed some light on the possible roles of the observed distribution for miRNAs. We found that unlike miRNA precursors, mirtronic precursors did not exhibit the characteristic distribution or enrichment for binding sites (supplementary data 5, figure 2), suggesting a possible regulatory role of the observed miRNA-specific enriched region in precursor transcription.

Such observed characteristic regions could have some regulatory impact on the transcription of precursors. We used some high-accuracy tools to predict transcription initiation potential, based on RNA-pol-II-dependent coding gene models. We analysed the precursor miRNA sequences with 1 Kb upstream sequences for transcription start site potentiality. Interestingly, in support of our previous findings, we found positive signals for TSS in the regions closer to several precursors (supplementary data 1; figures 3 and 4). As observed previously for distribution of binding sites, here, too, we did not find much difference between intergenic and intronic miRNAs, which showed almost identical distributions of TSS.

3.2 *Expression data correlation analysis between miRNAs and transcription factors suggests regulatory role of the preferred region*

As reported above, TFBS clustering was observed for the preferred region, which could be critical in transcriptional regulation of miRNAs. An expression-data-based correlation study was performed between each miRNA and the corresponding transcription factors which had binding sites present in the preferred region of the associated precursor miRNA. Binding sites for 157 different transcription factors were observed for 1030 miRNAs. 1001 unique miRNAs exhibited transcription factor binding sites in the enriched binding region for the given precursor miRNA sequences. However, the expression data for miRNAs analysis in mimiRNA was available for only 550 human miRNAs. For every reported miRNA, multiple TFBS were obtained in the enriched region. We measured the expression similarity correlation coefficient for each miRNA and its corresponding transcription factors whose binding sites were present in the preferred region. Finally, only those miRNA:TF associations were considered which had highest absolute correlation coefficient value. This way, one-to-one relationships between transcription factors and corresponding miRNAs were formed. However, for some transcription factors, too, expression and correlation data were not available. Therefore, instead of 550 correlations, we were able to get 379 correlations. Interestingly, 301
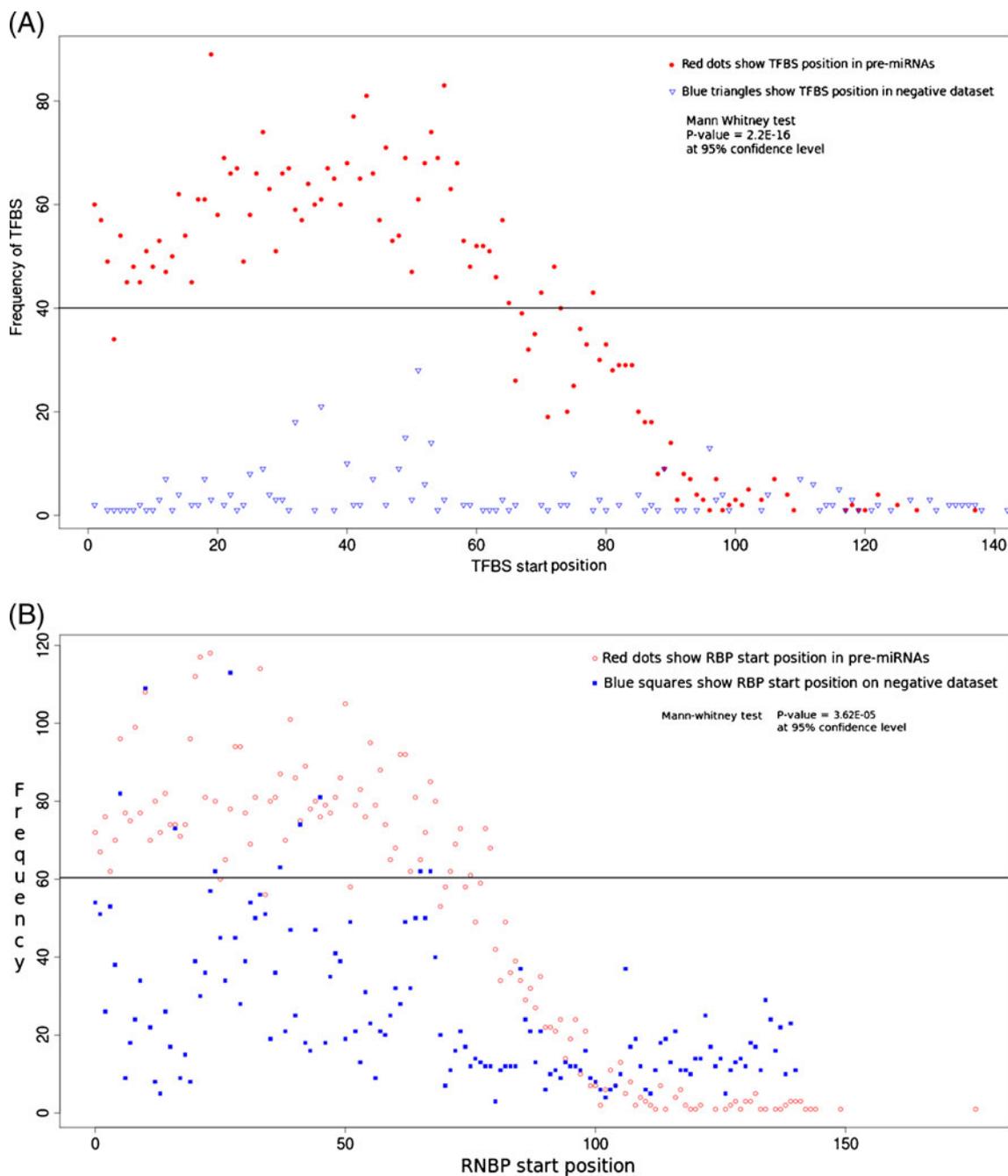
Figure 2. miRNA precursor sequences exhibit distinct distribution pattern for binding sites. A comparative study was performed between the human miRNA precursor sequences as the positive dataset and non-miRNA sequences as the negative dataset. (**A**) Distribution for TFBS and (**B**) distribution for RBP. The distribution was found to be higher as well as characteristically different for miRNAs with very high significance value (*P*-value << 0.001).

correlations out of 379 were positive. We categorized these 379 correlations into six expression categories, whose details are listed in table 1. The ratio of intronic *vs* intergenic miRNAs was almost similarly distributed in all of these five categories, suggesting no preference for any particular expression category either by intronic or intergenic

miRNAs. About 70 miRNA–TF relationships showed positive expression correlation value above 0.5. Supplementary data 1 shows the miRNA families associated with these transcription factors. Table 2 shows the miRNAs and associated TFs having highest expression correlations (*r*>0.5). Genomic coordinates of these pre-miRNAs, having the highest
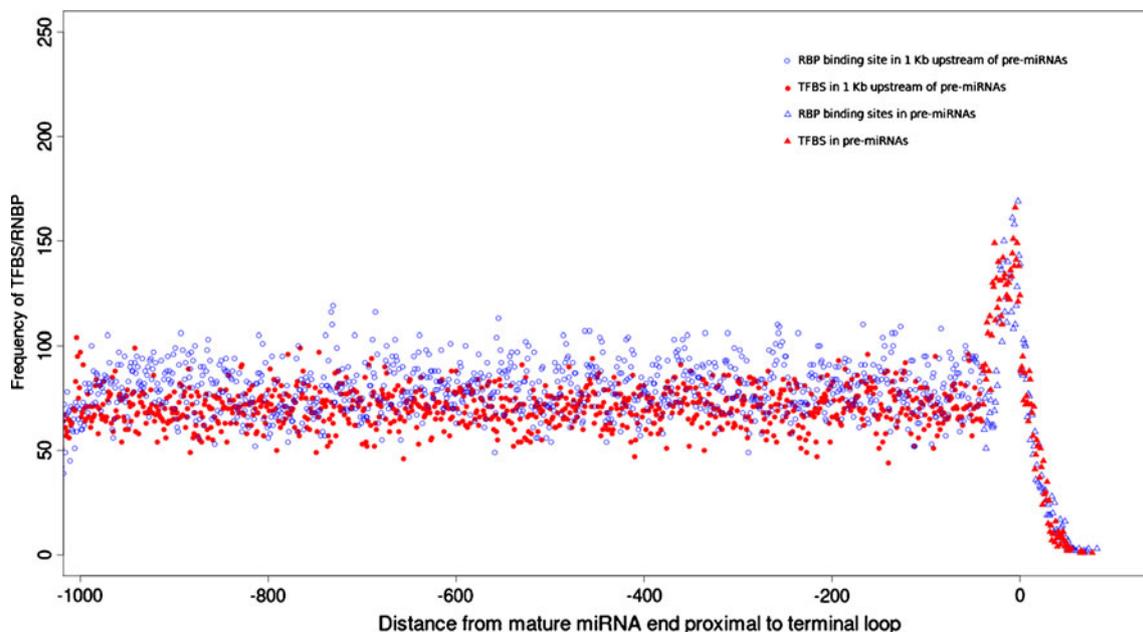
Figure 3. Distribution of regulatory sites in genomic regions containing the precursor miRNA sequences and their 5′ upstream 1 kb regions. The distribution remains almost uniform till the pre-miRNA region. A sudden increase in regulatory element frequency is observed in the precursor miRNA region, which attains sharp peak around the mature miRNA end closer to terminal loop region, suggesting enriched regulatory sites in the region which could be critical for regulation of miRNA biogenesis.

expression correlation with associated transcription factors in the enriched region, were retrieved from UCSC. Forty-eight miRNAs belonged to the intronic regions of various host genes. In order to find out if these miRNAs had preference for being co-transcribed with their host genes or not, we considered comparing the expression profiles of all of the intronic miRNAs with the expression profiles of their host genes. We measured expression correlation coefficients for all such miRNAs and associated host genes. Barring a few, most of them were either poorly or negatively correlated to their host genes. All these intronic miRNA, however, displayed high expression correlation with the transcription factors, which had binding sites in the enriched region of these miRNAs, suggesting transcriptional independence. This made us investigate further and we extended our study over these 70 highest correlated miRNA:TF associations through looking into occurance of composite regulatory modules and synergistic associations with other TFBS in their critical regions. Surprisingly, 60% of these pre-miRNAs, having high scoring associations with given transcription factors, exhibited significant existence of experimentally validated CRMs, within the preferred critical region in combination with the highly correlated transcription factors (*P*-value << 0.01). This means that most of the tanscription factors, which had bindng sites in the enriched regions of precursor miRNAs, and exhibit high expression correlation with

their host precursor miRNA, were present in these experimentally validated co-regulatory modules. Lmo2-GATA-1 module was observed as the most predominant one in this group (table 3).

3.3 *The preferred region houses binding slots for RNA-binding proteins, which may control specificity in miRNA biogenesis and processing after transcription*

Like the transcription factors, RBPs also exhibited positional preference and clustering in the preferred region of precursor miRNA sequences. A total of 954 unique miRNAs in association with very high scoring RBPs were present for this region. The sharpest tangents, exhibiting sudden drop in binding sites, were observed for distance positions −14 and +2, suggesting a critical role of the region falling in this interval. The highest number for binding in the preferred region was reported for MBNL-1, RBMX, yTHDC, PUM2, EIF4B, KHSRP, FUS and NONO. Following the same protocol as was done for transcription factor-miRNA expression data comparisons, expression correlations for 422 miRNA-RBP relationships were measured. ~60% RBP:miRNA correlations were positive. When categorized for the various intervals of positive correlations, only 32 relationships had positive correlations above the value of 0.5. Table 4 shows the various miRNA:RBP associations falling into this category.
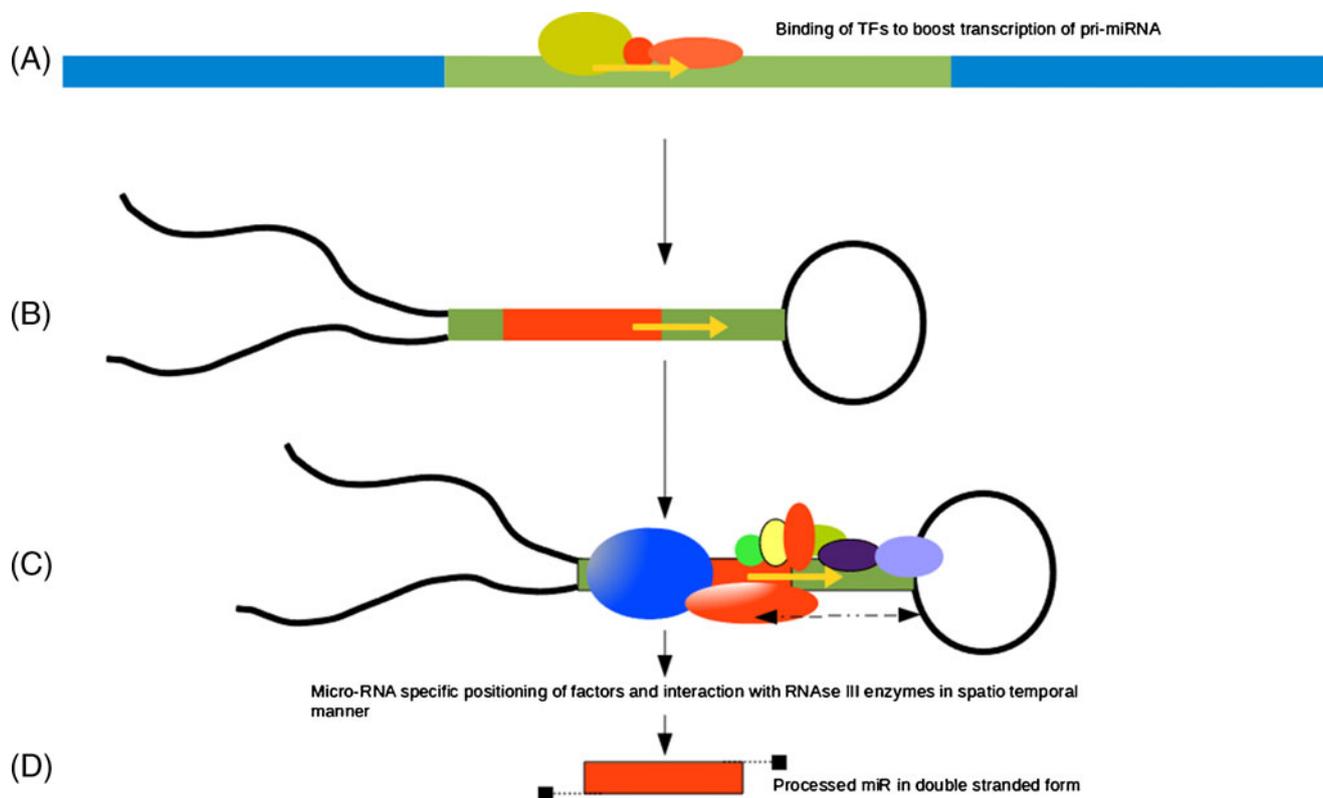
Figure 4. Critical region's possible significance in micro-RNA formation. (**A**) The transcription factors bind the critical region in the genomic sequence, in spatio temporal manner, to promote miRNA transcription. (**B**) Precursor miRNA sequence contains the critical region for binding of various RBPs. (**C**) RBPs bind the critical preferred region and establish interaction with RNA and RNAse III enzymes like Drosha-DGCR8 complex or Dicer. This could be also be a combinatorial process involving various RBPs in synergistic manner to influence binding of other binding factors including RBPs. (**D**) The processed ds miRNA.

Interestingly, most of the major associated RBPs in this group were associated with splicing/RNA-biogensis having multiple roles. Further investigation was done to find out the possible reasoning for splicing associated predominance. Genomic coordinates of all of the associated 32 miRNAs

Table 1. Characterization of miRNA TF co-expression according to correlation coefficient

Correlation coefficient r	Number of miRNA-TFBS associations
0–0.099	51
0.1–0.199	57
0.2–0.299	47
0.3–0.399	50
0.4–0.499	26
0.5–above	70

The table shows the distribution of miRNA-TF complex having positive expression correlation in the given range. The chi-square test shows the enrichment for the highest correlated category as significant one (P-value<0.001).

were located. Twenty-one of these miRNAs were from intronic regions. The highest number of such miRNAs were found to be associated with MBNL-1 and PABPC-1, which have been found to be associated with splicing and stability associated functions so far. Based upon family, factor and expression, we formed different groups/clusters (supplementary data 1). For positively associated top scoring correlations, MBNL-1, PUM2 and PABPC1 appeared as the most frequently occurring factors associated with many miRNAs. We further compared the distribution of various RBPs across negatively and positively correlated associations with miRNAs and found a few RBPs whose distributions were highly skewed between negatively and positively correlated groups. A2BP1, HNRNPA1, KHSRP, MBNL1, PABPC1 and PUM2 were the RBPS that exhibited enrichment in miRNA families having positive expression correlations. Among the negatively correlated miR:RBP associations for expression, we found RBMX as the predominant one, existing with ~20% (33 out of 167) of such instances. Fourteen miRNA:RBP associations exhibited highest negative correlations having values above 0.5. Six of them were associated with RBMX. Interestingly, for several of

Table 2. miRNAs and associated transcription Factors in enriched region having high expression correlation

TF	Function keyword	miRNA(correlation coefficient)
Hand-I/E47	Cell differentiation	hsa-mir-6210(0.713); hsa-mir-512-1(0.685); hsa-mir-512-2(0.685); hsa-mir-661(0.578);hsa-mir-301b(0.554)
GBP	Cell cycle and signalling	hsa-mir-632(0.803); hsa-mir-598(0.794); hsa-mir-589(0.792); hsa-mir-592(0.789); hsa-mir-302a(0.704); hsa-mir-650(0.596); hsa-mir-33b(0.578); hsa-mir-576(0.539); hsa-mir-877(0.513)
LMO2	Hemopoeisis	hsa-mir-101-1(0.929); hsa-mir-643(0.807); hsa-mir-363(0.773); hsa-mir-568(0.558); hsa-mir-580(0.542); hsa-mir-454(0.523)
SOX-9	Developmental	hsa-mir-374b(0.904);hsa-mir-588(0.843);hsa-mir-556(0.764); hsa-mir-586(0.579); hsa-mir-338(0.577)

The given transcription factors are those who were predominant in the highly correlated associations.

miRNA-RBP groups, all those miRNAs that shared a common RBP, binding them in the enriched region with high absolute expression profile correlations, also shared high positive correlation values for their expression profiles among themselves; strongly suggesting some common regulatory function of these RBPs with associated miRNAs, which appear to have some common role (supplementary data 4).

4. Discussion

For biogenesis of miRNA, two key proteins, Drosha and Dicer, have been identified as the key enzymes for a long time now (Hutvagner *et al.* 2001; Lee *et al.* 2003). A few recent reviews have discussed the degree of complexity in miRNA biogenesis in detail, suggesting the need of extensive work in this area while considering critical roles played by other factors besides Drosha and Dicer in miRNA biogenesis (Davis-Dusenbery and Hata 2010; Siomi and Siomi 2010; Slezak-Prochazka *et al.* 2010) . There are many alternative pathways of miRNA biogenesis as well as many other important components of miRNA biogensis that constitute various points of regulation in miRNA biogenesis (Cheloufi *et al.* 2010; Newman and Hammond 2010; Slezak-Prochazka *et al.* 2010). In this study, we have identified a preferred enriched region in the precursor miRNA sequences, which houses binding sites for many transcription factors as well as RBPs. This preferred region is centered around the mature miRNA region's end, proximal to the terminal loop region. The enriched region spans miRNA and some portion of the terminal loop. As the results above show, the highest density of such factors is found nearest to the mature miRNA region's end (reference point, as mentioned in methodology section) proximal to the terminal loop region (figure 1) with approximately symmetrical distribution around the peak. A sharp fall in transcription factor binding sites occur at positions −18 and +11, recording highest tangent values for the left side and right side of the plot, respectively. This region is most

significantly enriched for binding sites (*P*-values << 0.001) (table 5). Such a significant clustered arrangement of binding sites is supposed to be highly functional in the process of regulation and expression (Wasserman and Fickett 1998; Su *et al.* 2010). Almost the same pattern was observed for RBP binding sites, which recorded sharp fall at −14 bp upstream and +2 bp downstream. Region falling in this interval recorded lowest and significant *P*-values for enrichment (*P*-value << 0.001). We also found that this characteristic was common to miRNAs regardless of their genomic location as the similar distribution pattern was observed for them and analysed separately for intergenic and intronic miRNAs.

When compared with non-miRNA sequences as negative datasets, we found the above-mentioned enrichment property and distribution a significantly exclusive feature of precursor miRNA sequences, which could be useful in distinguishing a miRNA region from non-miRNAs (figure 2). As part of this study, we performed binding site analysis on precursor miRNA sequences along with their corresponding 1 kb upstream genomic sequences. It is believed that the upstream flanking region holds regulatory potential to promote transcription. However, we found an almost uniform distribution pattern of binding sites in the 1 kb upstream regions of the pre-miRNA regions till the beginning of the pre-miRNA region, which witnessed a steep rise with sharp peak centered around the mature miRNA region's end, proximal to the terminal loop start region (reference position) (figure 3). Performing this part of analysis served two purposes: (1) An additional negative dataset was gained through 1 kb upstream region sequences as the control to compare the distribution of binding sites with respect to the precursor miRNA sequences, besides the above-mentioned Wilcox test and (2) a contextual and broader comparison with associated upstream region could be done, showcasing the behavior of binding pattern distribution in a larger frame, suddenly peaking around the preferred critical region in pre-miRNA sequence regions. Such enrichment of binding sites in the preferred region could provide an array

Table 3. Composite regulatory module detected for highly correlated miRNA-transcription factor associations

miRNA	Composite regulatory elements	*P*-value/Reference
hsa-mir-539	IRF-PU1	2.70E−003
	ELF1-AP1	4.17E−003
	GATA1-Lmo2	[1]
hsa-let-7 g		
	GATA1-Lmo2	[1]
	ELF1-AP1	2.37E−002
hsa-mir-655		
	ELF1-AP1	7.03E−003
	AP1-ELF1	9.27E−003
	IRF-ETS	1.95E−002
	IRF-ETS	1.99E−002
	OCT-HMGIY	3.23E−002
	GATA1-Lmo2	[1]
hsa-mir-17		
	GATA1-Lmo2	[1]
hsa-mir-542		
	Lmo2-Hand1/E47	5.00E−004
	PU1-IRF	4.47E−003
hsa-mir-331		
	HNF3-HNF3	2.80E−003
	Hand1/E47	[2]
hsa-mir-599		
	CEBPA-ETS	1.29E−002
	IRF-ETS	3.72E−002
	GATA1-Lmo2	[1]
hsa-mir-409		
	Hand1/E47	[2]
hsa-mir-129-1		
	ETS-PEBP	6.22E−006
	MYB-ETS	4.12E−004
	Hand1/E47	[2]
hsa-mir-199b		
	IRF-ETS	4.80E−002
	GATA1-Lmo2	[1]
hsa-mir-620		
	OCT-HMGIY	4.05E−002
	GATA1-Lmo2	[1]
hsa-mir-589		
	Hand1/E47	[2]
hsa-mir-592		
	CEBPB-CREBP1CJUN	3.79E−004
	CEBPA-ETS	3.80E−002
	ATF-CREB	5.00E−004
hsa-mir-33b		
	USF-USF	5.00E−004
	HMGIY-OCT	8.48E−003
hsa-mir-621		
	ELF1-AP1	8.27E−004
hsa-mir-512-1		
	E47-E47	5.00E−004
hsa-mir-512-2		

Table 3. (continued)

miRNA	Composite regulatory elements	*P*-value/Reference
	E47-E47	5.00E−004
hsa-mir-661		
	Hand1/E47	[2]
hsa-mir-301b		
	Hand1/E47	[2]
	STAT-CEBP	2.57E−003
hsa-mir-101-1		
	GATA1-Lmo2	[1]
	IRF-ETS	3.44E−002
hsa-mir-643		
	ELF1-AP1	1.13E−002
	HNF3-NF1	1.44E−002
	HNF3-NF1	2.64E−002
	AP1-ELF1	2.67E−002
	GATA1-Lmo2	[1]
hsa-mir-363		
	GATA1-Lmo2	[1]
	HMGIY-OCT	1.18E−002
hsa-mir-568		
	OCT-HMGIY	2.90E−002
	HMGIY-OCT	3.28E−002
	GATA1-Lmo2	[1]
hsa-mir-580		
	NFAT-AP1	1.15E−004
	HMGIY-NFKB	1.34E−004
	AP1-NFAT	1.95E−004
	ELF1-AP1	2.07E−003
	AP1-ELF1	4.04E−003
	ELF1-AP1	4.49E−002
	GATA1-Lmo2	[1]
hsa-mir-378		
	RFX1-RFX1	8.00E−003
hsa-mir-139		
	USF-USF	5.00E−004
hsa-mir-99b		
	USF-USF	5.00E−004

About 45 pre-miRNAs exhibited highly significant and experimentally validated CRMs in the preferred binding zone. CRMs were detected using Transcompel and Cis-RED databases who archive the experimentally validated significant CRM associations. 42 out of 70 top scoring miRNAs, belonging to highest correlated goup for coexpression with associated transcription factors in the preferred regions,exhhibited significant and experimentally validated CRM, supporting the regulatory role of the preferred region. Each member of the CRM contains at least one TF, which was found to have positive expression correlation with the associated miRNA.

[1] Wadman *et al.* 1997

[2] Knöfler M, Meinhardt G, Bauer S, Loregger T, Vasicek R, Bloor DJ, Kimber SJ and Husslein P 2002 Human hand1 basic helix-loop-helix (bHLH) protein: extra-embryonic expression pattern, interaction partners and identification of its transcriptional repressor domains. *Biochem. J.* **361**

Table 4. RBPs having correlation coefficient above 0.5

RBP	Known role of RBP	Associated miRNA
a2bp1	Splicing/RNA biogenesis	Hsa-mir-668, hsa-mir-619, hsa-mir-596
EIF4B	Cap associated protein binding to RNA	Hsa-mir-483, hsa-mir-767, hsa-mir-520 h, hsa-mir-517b
MBNL1	Splicing/RNA biogenesis	Hsa-mir-600, hsa-mir-582, hsa-mir-661, hsa-mir-649, hsa-mir-581, hsa-mir-604
PABPC1	modulation of RNA activity and stability	Hsa-mir-363, hsa-mir-608, hsa-mir-567, hsa-mir-645, hsa-mir-643, hsa-mir-525, hsa-mir-548c, hsa-mir-651, hsa-mir-605, hsa-mir-586, hsa-mir-361, hsa-mir-610
NONO	Splicing	hsa-mir-331
Pum2	RNA stability	hsa-mir-15b, hsa-mir-568, hsa-mir-657, hsa-let-7 g
RBMX	Splicing/RNA-biogenesis	hsa-mir-590
YTHDC1	Splicing/RNA biogenesis	hsa-mir-518c

The associated miRNAs and RBPs binding in the preferred region exhibited expression correlation coefficient higher than 0.5.

of auxiliary factors to control miRNA transcription and post-transcriptional processing in highly a tissue- and family-specific manner. Besides this, it could also impart high conservation for miRNAs and nearby region which augers well with an earlier made observation on marked conservation of stem–loop regions in pre-miRNA sequences (Lai 2003). To add more confidence in these observations, comparative studies with mirtrons yielded interesting findings as the observed distribution pattern was not followed by mirtrons, the elements that are dependent upon host transcription process for formation (Berezikov *et al.* 2007) and produce miRNA-like elements after Drosha-independent processing. Such observations supported the possibility for regulatory role of the observed enriched regions, exclusive to miRNAs who have been reported to transcribe dually, i.e. co-transcribing with host as well as independently (Monteys *et al.* 2010). In a part of this study, we applied tools to measure the TSS potential in and around the precursor sequences, and found that for several sequences, positive signals were present near the miRNA regions, suggesting their potential to transcribe (supplementary data 5; supplementary figures 4 and 5). Like the distribution of regulatory regions, distribution of TSS was similar for intronic and intergenic miRNAs, indicating no visible impact of genomic location of miRNAs. However, these TSS potential measuring tools have been developed

exclusively for the RNA-pol-II-dependent system, where information of protein coding genes were used extensively, which could be a limiting factor for such observation. Such tools are destined to identify only those instances which somehow show similar trends towards the coding gene based features, like the spectrum of k-mers distribution in first exon and 5′ UTR regions. In that manner, there should be more such instances having transcribing potential.

The transcription factor binding sites in this region may make this region critical for transcription level regulation, while the RBP sites within this region may be critical during post-transcriptional processing of the precursor sequences. To corroborate these assumptions, we carried out expression-profile-based estimation of similarity between the miRNAs and the corresponding transcription factors and RBPs associated with the enriched binding region. About 80% of such transcription factors, having binding sites in the preferred region, showed positive correlations for expression profiles with respective associated miRNAs, a highly significant observation (*P*-value 2.2E−16). The highest number of miRNAs and associated transcription factors was observed for the category having expression correlation coefficient value above 0.5 (*P*-value <0.001; χ^2=20.627, DF=5). The transcription factors that appeared most in this category were SOX-9, Lmo-2, GBP, Hand-I/E47 and GATA families. Some of these associations may be critical to review. Transcription factor SOX-9 binds miRNAs hsa-miR-588, 556, 374b and 586 in the preferred region and all of them have high correlation coefficient for co-expression with SOX-9. SOX9 and the associated miRNA have been implicated in various carcinoma such as colorectal cancer and stand as critical components of regulation of cell cycle, differentiation and development (Jay *et al.* 2005). Transcription factors like GBP have been associated with signalling and proliferation and have been identified as a promising target in cancer therapy (Fellenberg *et al.* 2004; Guimarães *et al.* 2009). Similarly, Lmo-2 expression has been noted to be very high in pancreatic cancer cells and exhibits strong correlation to aggression of prostate cancer and tumour angiogenic conditions (Yamada *et al.* 2002; Ma *et al.* 2007; Nakata *et al.* 2009). Many of the these over-represented transcription factors have been identified to be associated with various forms of cancers. In this study, we have identified those miRNA who are associated with these factors in response and may be useful in understanding the cell behavior. These associations could be of therapeutic interest and should be investigated further. The close associations between transcription factors and associated pre-miRNAs having binding sites for them in the preferred region, majority of these associations showing positive expression correlation and the highest number of associations belonging to the highest expression correlation category, suggest a critical role of the enriched region in

Table 5. Significance of distribution of binding factors with respect to distance from 3′ end of mature miRNA region for the preferred region

Position	P-value (TFBS)	P-value (RBP)	Position	P-value (TFBS)	P-value (RBP)
−20	2.06E−07	0.000484	1	5.2E−34	3.55E−027
−19	7.32E−12	2.81E−015	2	9.8E−30	2.93E−043
−18	3.85E−12	4.11E−013	3	4.34E−28	7.49E−025
−17	4.48E−08	8.43E−007	4	2.79E−27	4.26E−034
−16	2.74E−13	4.45E−006	5	8.31E−21	4.51E−033
−15	9.53E−10	0.000212	6	1.76E−26	9.01E−026
−14	3.85E−12	6.60E−015	7	5.23E−23	2.63E−021
−13	8.88E−11	8.84E−011	8	9.65E−20	8.39E−017
−12	5.33E−10	4.45E−006	9	2.9E−22	7.04E−021
−11	5.21E−09	1.49E−006	10	1.24E−22	1.83E−010
−10	5.33E−10	0.000322	11	2.32E−18	3.74E−010
−9	1.06E−18	1.54E−014	12	1.4E−06	3.41E−017
−8	2.01E−12	4.45E−006	13	5.36E−13	4.36E−012
−7	2.59E−11	8.10E−014	14	8.78E−07	3.00E−009
−6	1.1E−17	1.49E−006	15	4.21E−05	4.45E−006
−5	5.36E−13	7.73E−008	16	0.000129	2.09E−005
−4	1.06E−18	3.00E−009	17	3.46E−06	2.59E−006
−3	4.37E−26	1.49E−006	18	1.9E−05	1.49E−006
−2	2.05E−15	9.12E−013	19	0.021208	2.09E−005
−1	9.65E−20	1.83E−010	20	0.005857	1.43E−007
0	5.4E−31	1.46E−030			

The observed statistics was done by applying Poisson distribution over the all distance positions observed for various pre-miRNAs and associated binding factors. As it can seen here, the preferred region scores high significance for enrichment, which increases with proximity to the 3′ end of the miRNA region and a bit of downstream region.

miRNA expression through influencing the process of transcription in a specific manner (figure 4).

Our findings received further support when we found that out of the 70 highest correlated miRNA–TF associations, as mentioned above, 48 miRNAs were from the intronic regions of various host genes and exhibited poor expression correlation, suggesting independent transcription of miRNAs. Some studies had suggested that intronic miRNAs are co-expressed with their host genes, and promoters of the host genes are involved in the transcription process of the miRNA region along with the host gene itself (Rodriguez *et al.* 2004). However, in our case, we did not find any such co-expression apparent, barring a few, for this group of miRNAs. For most of the intronic miRNAs belonging to this group, we observed either poorly correlated or negatively correlated host gene expression pattern. For such observation there could be two possible apparent reasons: Either (1) the resident miRNA is being co-transcribed by the host transcription system but step of maturation is being impeded after transcription, or (2) the resident microRNA is transcribing independent of the host system. However, if first case holds true, there should be at least a moderate expression profile similarity instead of negligible/poor

expression profile similarity. Recent studies have concluded well that usually resident miRNAs, when co-transcribed with the host genes, exhibit fair expression similarity with the host gene (Rodriguez *et al.* 2004; Monteys *et al.* 2010). Secondly, if processing impediment happens, it is independent of the mode of transcription and it could be apparent even when autonomous transcription is taking place. In that condition, the same intronic miRNAs that exhibited poor expression similarities with the host genes exhibited strong expression profile similarities with transcription factors that had strong binding regions in enriched regions of those precursor sequences and were found partners in majority of experimentally validated co-regulatory modules. Co-regulatory modules (CRMs) are experimentally known cases of response element pairs that facilitate the process of transcription together (Matys *et al.* 2006). When analysed for the existence of composite regulatory modules in the preferred enriched region of those 70 top scoring instances of pre-miRNA sequences and associated TF, we found statistically significant (P-value $\ll$ 0.01) and experimentally validated CRMs in a majority of these sequences, supporting the theory that the preferred region could be important for regulation of transcription. Lmo-2-GATA-1 associations

appeared as the most predominant ones among these regulatory modules. GATA-1 is a versatile transcription factor whose activity has been noted as highly context specific (Wozniak *et al.* 2007). In turn, Lmo-2 and GATA-1 association has been proven to be highly critical for development, where Lmo-2 preferentially interacts with GATA-1 through its LIM2 domain, which together cause active transcription (Wadman *et al.* 1997). Therefore, the second condition supporting independent transcription capacity is favoured. The remaining miRNAs were from intergenic regions which mostly hold independent transcription units in themselves. These observations are in concordance with observations made by several groups who reported the self-transcription of many intronic miRNAs as well as the dual mode of transcription where a miRNA could be transcribed with the host as well as independently (Isik *et al.* 2010; Monteys *et al.* 2010; Gao *et al.* 2010). Some of these groups reported that, contrary to the prevalent belief that intronic miRNAs usually get transcribed by the host gene transcription machinery, they found that at least one-third of intronic pre-miRNAs were transcribed independent of the host genes. Recently studies by Gao *et al.* (2010) reported that the miRNA precursor sequences can transcribe independent of external promoter elements and hold an innate capacity to transcribe themselves. The authors carried out construct-dependent experiments, with and without promoter sequences, containing precursor miRNAs. They found that precursor miRNAs, in the absence of any upstream promoters, were able to transcribe autonomously. In a series of well-executed experiments, this group concluded that precursor miRNAs can undergo transcription independently, which could be tissue specific, depending upon the kind of factors provided to them, which in turn could bind to the precursor miRNAs sequence and promote self-sufficient independent transcription.

The findings from this study may provide a good reasoning for the observed capacity of pre-miRNAs to transcribe independent of host genes or support the process of transcription. The preferred critical region may be useful in providing transcriptional potential to pre-miRNA on similar lines as has been observed for other already known autonomous elements like Alu elements. Shankar *et al.* (2004) gave a complete propagation model for Alu elements, which internally host an array of TFBS, rendering themselves transcriptionally active, sponsoring as well as autonomous. The existence of such regions enriched in regulatory elements has been found to have impact on transcription of even host genes, despite them being intronic (Bornstein *et al.* 1987; Seshasayee *et al.* 2000; Coulon *et al.* 2010). In light of the recent development with emergence of pervasive model of transcription, which states that almost all genomic locations have RNA polymerase binding capacity and there is enough evidence of independent transcription

being carried out by intronic regions too (Berretta & Morillon. 2009; Dinger *et al.* 2009), the findings made in this study gather apt relevance.

Very recently, few groups suggested that the process of miRNA formation may be regulated critically by other factors besides Drosha and Dicer (Cheloufi *et al.* 2010; Liu *et al.* 2008; Stark *et al.* 2008; Haase *et al.* 2005). RBPs have emerged as the guiding factor for RNA fate, taking over control during the process of transcription itself, and are involved in splicing, nuclear export, post-transcriptional processing, stability and degradation, localization and transport (Keene, 2001). Our present study with RBPs associated with miRNAs suggests for complex process of miRNA formation and existence, where RBPs could provide auxiliary apparatus for stage-, tissue- and miRNA-specific processing, localization and stability. 60% of miRNA:RBP unique associations in the preferred region were positively correlated for the expression data (*P*-value = 2.143E−05). Despite this skew, although lesser than the one observed for transcription factors, very few RBP:miRNA associations (just 32) showed positive correlation above 0.5, while largest number exhibited lowest level of correlation (90). This suggests that despite having binding sites for many RBPs in most miRNAs, limited number of RBP-miR associations show strong positive correlations (supplementary data 2). For many RBPs, we found high-confidence binding regions but poor expression correlations. In the present study, we used expression data of mature miRNAs. Expression data of a mature miRNA could be vastly independent from the expression data for prE−and primary miRNAs, pertaining to the influence of various associated factors (Fukuda *et al.* 2007). One explanation for such observations could be the requirement of combined regulation by more than one factor, like composite regulatory modules, which are involved in many cases of transcriptional regulation by transcription factors in a concerted manner or highly tissue-specific regulation of the particular miRNA by the associated factor. Combinatorial impact of RBPs on RNA processing and trafficking has been acknowledged to work significantly in a grouped manner to control the RNA level in highly specific manner. Also, large number of RBPs may be involved in the trafficking and localization process of miRNAs, as has been observed for several RNAs. In such situations, expression data may not be sufficient to draw a conclusion. A transcript can belong to more than one RNP complex, which renders it regulated differentially (Keene 2007). In such situations also, information from gene expression correlation measurement cannot be of much use.

At this point in the discussion, it would be worthwhile remembering those 32 highest correlated instances of RBPs and miRNAs (table 4) as well as those predominantly

associated with positively correlated groups of miRNAs for the preferred region (A2BP1, HNRNPA1, KHSRP, MBNL1, PABPC1 and PUM2). PUM2, PABPC1, A2BP1 and MBNL1 have also been found to be present among those 32 instances having highest correlation coefficient values. The majority of the associated RBPs in this category belonged to splicing-mediated RNA biogenesis. When investigated further, we found that majority of miRNAs in these 32 associations were intronic, going in line with some previous studies which suggest association of splicing factors with miRNA processing (Shiohama *et al.* 2007). Interestingly, some recent studies have found several splicing-associated RBPs involved with precursor miRNAs and their processing. SF2/ASF1 has been a known splicing factor but were recently found associated with a pre-miRNA, promoting Drosha-mediated miRNA processing (Wu *et al.* 2010). kHSRP is another such case, first characterized as a 75 kDa component of a multi-protein splicing complex, carrying out splicing. However, Trabucchi *et al.* (2009) identified Drosha-Dicer-mediated miRNA maturation as one more function carried out by kHSRP. hnRNP A1 has been a long known factor involved in alternate splicing (Mayeda and Krainer 1992). However, recent work found hnRNP A1 regulating pri-miRNA processing by changing its secondary structure, creating a more accessible site for cleavage by Drosha (Michlewski *et al.* 2008). MBNL1 (Goers *et al.* 2010) and A2BP1 (Auweter *et al.* 2006) regulate splicing/RNA biogenesis and their enrichment in the critical region may hint towards use of some mechanisms commonly involved in splicing as well as biogenesis of associated miRNAs. MBNL-1 has inherent affinity towards a stable hairpin structure in order to interact with RNAs and assist in its processing (Yuan *et al.* 2007). In our case, the found pre-miRNAs rich in MBNL-1 binding and strongly correlated in expression with MBNL-1, which could provide hairpin structures ably. Therefore, for the observed 32 RBP-miRNA associations, predominance of splicing-associated RBPs hint towards their role in miRNA processing much like the splicing factors found involved in miRNA processing, which could modulate the structure of pre-miRNA for processing. Other factors like PABPC1 interact with an RNA through their A-rich region and have impact on the stability of the RNA molecule. Pum2, a member of a group of highly pleiotropic RBPs, is a very likely candidate for RNA stability as well as specific transportation (Gerber *et al.* 2004; Keene 2007; Vesseya *et al.* 2010). Pumilios have been conferred with many functions. One of them has been their association with RNA secondary structures, binding to hairpin-stem structures and making them available for miRNA targeting (Kedde *et al.* 2010). Similarly, association of Pumilio with pre-miRNAs may bring structural changes required for control of miRNA processing. One recent study has found them assisting let-7 miRNAs in finding their targets (Nolde *et al.*

2007). This is a kind of mechanism adopted by Ago proteins to guide miRNAs to targets. In our study, Pum-2 was found associated with let-7 miRNA, which shares high expression correlation also. Although Pum-2 is a versatile RBP, it is difficult for us to tell the exact function of this association in the present study. However, as mentioned above, and considering its already determined functions as transporter of RNA in specific manner, it could be assisting in let-7 transport to the specific target or helps prE−let-7 attain the desirable conformation as per the need. Another RBP, RBMX, a kind of hnRNP, was found to be predominantly associated with miRNAs having negative expression correlation with them. Even the group with highest negative expression correlation had predominance of RBMX (hsa-mir-589,hsa-mir-588,hsa-mir-488,hsa-mir-665,hsa-mir-146b,hsa-mir-591). As discussed above, hnRNPs are among the most pleiotropic RBPs. Also, very little is known about the RBP function of RBMX, which was previously better known as a transcription factor. However, one recent study has found RBMX's preference for hairpin stem–loop structures in RNAs (Kanhoush *et al.* 2009).

In the concluding part of this study with RBPs, we came across a very encouraging finding, obeying the Regulon Model exactly as proposed by Keene (2007). The identified miRNA–RBP associations, having high absolute expression correlations, were classified further into groups, where all miRNAs sharing a common RBP were classified into the same group identified by the associated RBP. The expression profiles of every miRNA belonging to same groups were compared, and expression correlation was measured. Interestingly, for many RBP groups, we found strong co-expression among the member miRNAs, clearly approving our findings as well as suggesting some common biological roles being played by all of these miRNAs, which are centrally commanded by the associated RBPs (supplementary data 4). This is in exact agreement with the Regulon Model proposed by Keene (2007). In general, RBPs have emerged as multifunctional proteins that cannot be associated with one single function alone as they participate in several aspects of RNA biochemistry (Licatalosi and Darnell 2010). However, with the current study, these above-mentioned factors, and their associated miRNAs for the enriched region, emerge as highly potential candidates in which RBPs may be associated by contributing in the processing, stability, movement or maturation of associated miRNAs.

The process of miRNA biogenesis seems to have multiple-level checkpoints where the enriched region in the pre-miRNA sequences stand as a critical regulatory candidate, housing TFBS to facilitate transcription, while RBPs binding there carry on post-transcriptional processes (figure 4). The present study provides a reasonable insight and a certain direction from findings made here, providing a template for experimental studies in order to

understand the process of miRNA formation and their cellular behaviour precisely.

References

Abeel A, Peer YV and Saeys Y 2009 Toward a gold standard for promoter prediction evaluation. *Bioinformatics* **25** i313–i320

Aguda BD, Kim Y, Piper-Hunter MG, Friedman A and Marsh CB 2008 miRNA regulation of a cancer network: consequences of the feedback loops involving miR-17–92, E2F, and Myc. *Proc. Natl. Acad. Sci. USA* **105** 19678–19683

Auweter SD, Fasan R, Reymond L, Underwood JG, Black DL, Pitsch S and Allain FH-T 2006 Molecular basis of RNA recognition by the human alternative splicing factor Fox-1. *EMBO J.* **25** 163–173

Berezikov E, Chung W, Wills J, Cuppen E and Lai EC 2007 Mammalian mirtron genes. *Mol. Cell.* **28** 328–336

Berretta J and Morillon A 2009 Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep.* **10** 973–982

Borchert GM, Lanier W and Davidson BL 2006 RNA polymerase III transcribes human miRNAs. *Nat. Struct. Mol. Biol.* **13** 1097–1101

Bornstein P, Mckay J, Morishima JK, Devarayalu S and Gelinas RE 1987 Regulatory elements in the first intron contribute to transcriptional control of the human al(I) collagen gene. *PNAS* **84** 8869–8873

Bushati N and Cohen SM 2007 miRNA functions. *Annu. Rev. Cell Dev. Biol.* **23** 175–205

Chang TC, Wentzel EA, Kent OA, Ramachandran K, Mullendore M, Lee KH, Feldmann G, Yamakuchi M, *et al.* 2007 Trans-activation of miR-34a by p53 broadly influences gene expression and promotes apoptosis. *Mol. Cell* **26** 745–752

Cheloufi S, Dos Santos CO, Chong MM and Hannon GJ 2010 A Dicer-independent miRNA biogenesis pathway that requires Ago catalysis. *Nature (London)* **465** 584–589

Cook BK, Kazan H, Zuberi K, Morris Q and Hughes TR 2010 RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.* **39** D301–D308

Coulon V, Chebli K, Cavelier P and Blanchard JM 2010 A novel mouse c-fos intronic promoter that responds to CREB and AP-1 is developmentally regulated in vivo. *PLoS One* **5** e11235

Davis BN, Hilyard AC, Nguyen PH, Lagna G and Hata A 2010 Smad proteins bind a conserved RNA sequence to promote microRNA maturation by Drosha. *Mol. Cell* **39** 373–384

Davis-Dusenbery BN and Hata A 2010 Mechanism of control of miRNA biogenesis. *J. Biochem.* **148** 381–392

Dinger ME, Amaral PP, Mercer TR and Mattick JS 2009 Pervasive transcription of the exukaryotic genome: functional indices and conceptual implications. *Brief. Bioinformatics* **8** 407–423

Down TA and Hubbard TJP 2002 Computational detection of location of transcription start sites in mammalian genomic DNA. *Genome Res.* **12** 458–461

Fellenberg F, Hartmann TB, Dummer R, Usener D, Schadendorf D and Eichmüller S 2004 GBP-5 splicing variants: New guanylate−binding proteins with tumor-associated expression and antigenicity. *J. Invest. Dermatol.* **122** 1510–1517

Filipowicz W, Bhattacharya SN and Sonenberg N 2008 Mechanisms of post-transcriptional regulation by miRNAs: are the answers in sight? *Nat. Rev. Genet.* **9** 102–114

Fukuda T, Yamagata K, Fujiyama S, Matsumoto T, Koshida L, Yoshimura K, Mihara M, Naitou M, *et al.* 2007 DEAD-box RNA helicase subunits of the Drosha complex are required for processing of rRNA and a subset of miRNAs. *Nat. Cell Biol.* **9** 604–611

Gao J, Zhang Y, Li M, Tucker LD, Machan JT, Quesenberry P, Rigoutsos I and Ramratnam B 2010 Atypical transcription of miRNA gene fragments. *Nucleic Acids Res.* **38** 2775–2787

Gerber AP, Herschlag D and Brown OP 2004 Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeasts. *PLoS Biol.* **2** E79

Goers ES, Purcell J, Voelker RB, Gates DP and Berglund JA 2010 MBNL1 binds GC motifs embedded in pyrimidines to regulate alternative splicing. *Nucleic Acids Res.* **38** 2467–2484

Griffiths-Jones S, Saini HK and Dongen SV 2008 miRBase: tools for miRNA genomics. *Nucleic Acids Res.* **36** D154–D158

Guil S and Cáceres JF 2007 The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nat. Struct. Mol. Biol.* **14** 591–596

Guimarães DP, Oliveira IM, de Moraes E, Paiva GR, Souza DM, Barnas C, Olmedo DB, Pinto CE, *et al.* 2009 Interferon-inducible guanylate binding protein (GBP)-2: a novel p53-regulated tumor marker in esophageal squamous cell carcinomas. *Int. J. Cancer* **124** 272–279

Haase AD, Jaskiewicz L, Zhang H, Lainé S, Sack R, Gatignol A and Filipowicz W 2005 TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. *EMBO Rep.* **6** 961–967

Han J, Lee Y, Yeom K-H, Kim Y-H, Jin H and Kim VN 2004 The Drosha–DGCR8 complex in primary miRNA processing. *Genes Dev.* **18** 3016–3027

Higo K, Ugawa Y, Iwamoto M and Korenaga T 1999 Plant cis-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Res.* **27** 297–300

Hofacker IL 2003 Vienna RNA secondary structure server. *Nucleic Acids Res.* **31** 3429–3431

Hutvagner G, McLachlan J, Pasquinelli AE, Bálint E, Tuschl T and Zamore PD 2001 A cellular function for the RNA-interference

enzyme dicer in the maturation of the let-7 small temporal RNA. *Science* **293** 834–838

Isik M, Korswagen HC and Berezikov E 2010 Expression patterns of intronic microRNAs in *Caenorhabditis elegans. Silence* **1** 5

Jay P, Berta P and Blache P 2005 Expression of the carcinoembryonic antigen gene is inhibited by SOX9 in human colon carcinoma cells. *Cancer Res.* **65** 2193–2198

Johnson SM, Lin SY and Slack FJ 2003 The time of appearance of the *C. elegans* let-7 miRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. *Dev. Biol.* **259** 364–79

Kanhoush R, Beenders B, Perrin C, Moreau J, Bellini M and Penrad-Mobayed M 2009 Novel domains in the hnRNP G/RBMX protein with distinct roles in RNA binding and targeting nascent transcripts. *Nucleus* **1** 109–122

Kaye JA, Rose NC, Goldsworthy B, Goga A and L'Etoile ND 2009 A 3′UTR pumilio-binding element directs translational activation in olfactory sensory neurons. *Neuron* **61** 57–70

Kedde M, van Kouwenhove M, Zwart W, Oude Vrielink JA, Elkon R and Agami R 2010 A Pumilio-induced RNA structure switch in p27-3′ UTR controls miR-221 and miR-222 accessibility. *Nat. Cell Biol.* **12** 1014–1020

Keene JD 2001 Ribonucleoprotein infrastructure regulating the inflow of genetic information between the genome and proteome. *Proc. Natl. Acad. Sci. USA* **98** 7018–7024

Keene JD 2007 RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.* **8** 533–543

Lai EC 2003 miRNAs: Runts of the genome assert themselves. *Curr. Biol.* **13** R925–R936

Lee RC, Feinbaum RL and Ambros V 1993 The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75** 843–854

Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, Lee J, Provost P, Rådmark O, Kim S and Kim VN 2003 The nuclear RNase III Drosha initiates miRNA processing. *Nature (London)* **425** 415–419

Licatalosi DD and Darnell RB 2010 RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.* **11** 75–87

Liu X, Fortin K and Mourelatos Z 2008 miRNAs: Biogenesis and Molecular Functions. *Brain Pathol.* **18** 113–121

Lee Y, Kim M, Han J, Yeom KH, Lee S, Baek SH and Kim VN 2004 miRNA genes are transcribed by RNA polymerase II. *EMBO J.* **23** 4051–4060

Ma S, Guan XY, Beh PSL, Wong KY, Chan YP, Yuen HF, Vielkind J and Chan KW 2007 The significance of LMO2 expression in the progression of prostate cancer. *J. Pathol.* **211** 278–285.

Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, *et al.* 2006 TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34** D108–D110

Mayeda A and Krainer AR 1992 Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. *Cell* **68** 365–375

Michlewski G, Guil S, Semple CA and Cáceres JF 2008 Posttranscriptional regulation of miRNAs harboring conserved terminal loops. *Mol. Cell* **32** 383–393

Monteys MA, Spengler RM, Wan J, Tecedor L, Lennox KA, Xing Y and Davidson BL 2010 Structure and activity of putative intronic miRNA promoters. *RNA* **16** 495–505

Nakata K, Ohuchida K, Nagai E, Hayashi A, Miyasaka Y, Kayashima T, Yu J, Aishima S, *et al.* 2009 LMO2 is a novel predictive marker for a better prognosis in pancreaticcancer. *Neoplasia* **11** 712–719

Newman MA, Thomson JM and Hammond SM 2008 Lin-28 interaction with the Let-7 precursor loop mediates regulated miRNA processing. *RNA* **14** 1539–1549

Newman MA and Hammond SM 2010 Emerging paradigms of regulated miRNA processing. *Genes Dev.* **24** 1086–1092

Nolde MJ, Saka N, Reinert KL and Slack FJ 2007 The Caenorhabditis elegans pumilio homolog, puf-9, is required for the 3′UTR-mediated repression of the let-7 microRNA target gene, hbl-1. *Dev Biol.* **305** 551–563

Ritchie W, Flamant S and Rasko JE 2009 mimiRNA: a miRNA expression profiler and classification resource designed to identify functional correlations between miRNAs and their targets. *Bioinformatics* **26** 223–227

Robertson G, Bilenky M, Lin K, He A, Yuen W, Dagpinar M, Varhol R, Teague K, *et al.* 2006 cisRED: a database system for genomE−scale computational discovery of regulatory elements. *Nucleic Acids Res.* **34** D68–D73

Rodriguez A, Griffiths-Jones S, Ashurst JL and Bradley A 2004 Identification of mammalian miRNA host genes and transcription units. *Genome Res.* **14** 1902–1910

Ruby GJ, Calvin HJ and Bartel DP 2007 Intronic microRNA precursors that bypass Drosha processing. *Nature (London)* **448** 83–87

Seshasayee D, Geiger JN, Gaines P and Wojchowski DM 2000 Intron 1 elements promote erythroid-specific GATA-1 gene expression. *J. Biol. Chem.* **275** 22969–22977

Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein MJ, Tuschl T, Nimwegen EV and Zavolan M 2005 Identification of clustered miRNAs using an ab initio prediction method. *BMC Bioinformatics* **6** 267

Shalgi R, Leiber D, Oren M and Pilpel Y 2007 Global and local architecture of the mammalian miRNA-transcription factor regulatory network. *PLoS Comput. Biol.* **3** e131

Shiohama A, Sasaki T, Noda S, Minoshima S and Shimizu N 2007 Nucleolar localization of DGCR8 and identification of eleven DGCR8-associated proteins. *Exp. Cell Res.* **313** 4196–4207

Siomi H and Siomi MC 2010 Posttranscriptional regulation of miRNA biogenesis in animals. *Mol. Cell* **38** 323–332

Slezak-Prochazka I, Durmus S, Kroesen B-J and Berg A 2010 miRNAs, macrocontrol: Regulation of miRNA processing. *RNA* **16** 1087–1095

Shankar R, Grover D, Brahmachari SK and Mukerji M 2004 Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements. *BMC Evol. Biol.* **4** 37

Sonnenburg S, Zien A and Ratsch G 2006 ARTS: accurate recognition of transcription starts in human. *Bioinformatics* **22** e472–e480

Stark KL, Xu B, Bagchi A, Lai W-S, Lui H, Hsu R, Wan X, Pavlidis P, Mills AA, Karayiorgou M and Gogos JA 2008 Altered brain

miRNA biogenesis contributes to phenotypic deficits in a 22q11-deletion mouse model. *Nat. Genet.* **40** 751–760

Stefani G and Slack FJ 2008 Small non-coding RNAs in animal development. *Nat. Rev. Mol. Cell Biol.* **9** 219–230

Su N, Wang Y, Qian M and Deng M 2010 Combinatorial regulation of transcription factors and miRNAs. *BMC Syst. Biol.* **4** 150

Tarasov V, Jung P, Verdoodt B, Lodygin D, Epanchintsev A, Menssen A, Meister G and Hermeking H 2007 Differential regulation of microRNAs by p53 revealed by massively parallel sequencing: miR-34a is a p53 target that induces apoptosis and G1-arrest. *Cell Cycle* **6** 1586–1593

Trabucchi M, Briata P, Garcia-Mayoral M, Haase AD, Filipowicz W, Ramos A, Gherzi R and Rosenfeld MG 2009 The RNA-binding Protein KSRP Promotes the Biogenesis of a Subset of miRNAs. *Nature (London)* **459** 1010–1014

Vesseya JP, Schoderboeck L, Gingl E, Luzi E, Reifler J, Di Leva F, Karra D, Thomas S, Keibler MA and Macchi P 2010 Mammalian Pumilio 2 regulates dendrite morphogenesis and synaptic function. *PNAS* **107** 3222–3227

Wadman IA, Osada H, Grütz GG, Agulnick AD, Westphal H, Forster A and Rabbitts TH 1997 The LIM-only protein Lmo2 is a bridging molecule assembling an erythroid, DNA-binding complex which includes the TAL1, E47, GATA-1 and Ldb1/NLI proteins. *EMBO J.* **16** 3145–3157

Wasserman WW and Fickett JW 1998 Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278** 167–181

Wickens M, Bernstein DS, Kimble J and Parker R 2002 A PUF family portrait: 3′UTR regulation as a way of life. *Trends Genet.* **18** 150–157

Wingender E, Dietze P, Karas H and Knüppel R 1996 TRANS-FAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* **24** 238–241

Wozniak RJ, Boyer ME, Grass JA, Lee Y and Bresnick EH 2007 Context-dependent GATA factor function: combinatorial requirements for transcriptional control in hematopoietic and endothelial cells. *J. Biol. Chem.* **282** 14665–14674

Wu H, Sun S, Tu K, Gao Y, Xie B, Krainer AR and Zhu J 2010 A splicing-independent function of SF2/ASF in microRNA processing. *Mol. Cell* **38** 67–77

Yamada Y, Pannell R, Forster A and Rabbitts TH 2002 The LIM-domain protein Lmo2 is a key regulator of tumour angiogenesis: a new anti-angiogenesis drug target. *Oncogene* **21** 1309–1315

Yousef M, Nebozhyn M, Shatkay H, Kanterakis S, Showe LC and Showe MK 2006 Combining multi-species genomic data for miRNA identification using a Naïve Bayes classifier. *Bioinformatics* **22** 1325–1334

Yuan Y, Compton SA, Sobczak K, Stenberg MG, Thornton CA, Griffith JD and Swanson MS 2007 Muscleblind-like 1 interacts with RNA hairpins in splicing target and pathogenic RNAs. *Nucleic Acids Res.* **35** 5474–5486

Zhang B, Pan X, Cobb GP and Anderson TA 2006 Plant miRNA: a small regulatory molecule with big impact. *Dev. Biol.* **289** 3–16

Corresponding editor: Reiner A Veitia