Songklanakarin J. Sci. Technol.
35 (2), 143-148, Mar. - Apr. 2013

SJST SONGKLANAKARIN JOURNAL OF SCIENCE AND TECHNOLOGY
http://www.sjst.psu.ac.th

Original Article

Computational identification of *Penaeus monodon* microRNA genes and their targets

Pitipol Meemak[1]*, Amornrat Phongdara[1], Wilaiwan Chotigeat[1], and Martti T. Tammi[1,2]

[1] *Department of Molecular Biotechnology and Bioinformatics, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, 90112 Thailand.*

[2] *Centre for Research in Biotechnology for Agriculture, Institute of Biological Science, University of Malaya, Kuala Lumpur, Malaysia.*

Received 15 June 2012; Accepted 25 January 2013

Abstract

MicroRNAs (miRNAs) are a distinct class of small non-coding RNAs, ~22 nt long, found in a wide variety of organisms. They play important regulatory roles by silencing gene activities at the post-transcriptional level. In this work, we developed a computational workflow to identify conserved miRNA genes in the 10,536 unique *Penaeus monodon* expressed sequence tags (ESTs). After removing all simple repeats and coding regions in the ESTs, the workflow uses both the conservation of miRNA sequences and several filters obtained from pre-miRNA secondary structure properties to identify conserved miRNAs. Finally, we discovered six potential conserved miRNA genes such as mir-4152, mir-466k, miR-32*, lin-4, mir-1346 and mir-4310.

Keywords: *Penaeus monodon*, miRNAs, MicroRNAs, expressed sequence tags, ESTs

1. Introduction

miRNAs are a class of small ~22 nt non-coding RNAs that are widely expressed in both plants and animals (Ambros, 2004; Bartel, 2004). In animals, they negatively regulate gene expression at the post-transcriptional level via sequence complementarity to the 3′ un-translated regions (3′ UTRs) of mRNAs (Bartel 2004). miRNAs have the ability to regulate many vital biological processes including cell proliferation, apoptosis, stem cell self renewal, differentiation, metabolism, organ development, developmental timing and tumor metastasis (Bartel, 2004; Williams, 2008; Huang *et al.,* 2011). Hence, many researches focused on the discovery of novel miRNAs to increase the understanding of their physiological functions. There are many approaches to identify miRNAs in various organisms. The first miRNAs, lin-4 and let-7, were identified in *Caenorhabditis elegans* using a genetic screening technology (Lee *et al.,* 1993; Wightman *et al.,* 1993). However, this method was limited by its expense and time consumption (Lai, 2003). Later, direct cloning and sequencing of short RNA molecules have been successfully applied to determine a number of miRNAs from animals and plant (Lagos-Quintana *et al.,* 2001; Reinhart *et al.,* 2002). However, the molecular cloning was limited to find miRNAs by highly constrained tissue- and time-specific expression patterns, presence of degradation products from mRNAs, and other non-coding RNAs (Lai, 2003; Lim *et al.,* 2003). To overcome these problems, computational technique was proposed to predict the potential miRNAs. In animals, the computational approaches identified miRNA genes by using known characteristics of miRNAs such as formation of hairpin loop secondary structure with a minimun folding free energy (Ambros *et al.,* 2003), the presence of mature miRNAs in the stem and not in the loop of the secondary structure, and high evolutionary conservation of mature miRNAs from species to species (Lai, 2003; Lim *et al.,* 2003). Recently, the most of

* Corresponding author.
Email address: pitipol.m@psu.ac.th

miRNAs in the database were identified by computational approaches and were subsequently verified by molecular techniques such as Northern blotting, PCR, and 5′ rapid amplification of cDNA ends (5′ RACE) (Griffiths-Jones, 2004; Griffiths-Jones *et al.,* 2006).

High-throughput sequencing technology has provided a mechanism to gain a genomics insight into species in the absence of a complete genome sequence by generating Expressed Sequence Tag (ESTs) collections. To date, the largest collection of ESTs from *Penaeus monodon* consists of 10,536 unique EST sequences from 37 cDNA tissue libraries (Tassanakajon *et al.,* 2006). An increasing number of miRNAs was reported and registered in the Wellcome Trust Sanger Institute miRBase (Griffiths-Jones, 2006; Griffiths-Jones *et al.,* 2006; Griffiths-Jones *et al.,* 2008; Griffiths-Jones, 2010). For example, 1,424 miRNAs have been reported from humans, 358 are from zebrafish, 238 from fruit fly and 612 from nematode. Recently, several miRNAs were identified from *Marsupenaeus japonicas* by using high-throughput sequencing of small RNAs (Huang *et al.*, 2012, Ruan *et al.*, 2011). Most of them were differentially expressed in response to white spot syndrome virus infection. Therefore, these miRNAs might play important roles in the *Marsupenaeus japonicas* immune system. Because of the high conservation of miRNAs, information regarding miRNAs and their targets in *Penaeus monodon* may contribute to an understanding of their role in gene regulatory networks across organisms.

Hence, this work is focused on how to identify *Penaeus monodon* miRNA genes in all the available EST sequences by using a computational screening approach. Utilizing the fact that miRNAs are highly conserved among the species (Weber, 2005), we developed a computational screening workflow to identify the miRNA orthologs. The use of EST sequence information will ensure that pre-miRNA candidates are expressed. Similar work has been successfully performed on human ESTs (Li *et al.,* 2006). In case of our study, it is not only provide a starting point for other further research on miRNAs, but also resulting in an improvement of understanding their functional roles in the *Penaeus monodon* genome.

2. Materials and Methods

2.1 Sequences data

All available miRNAs were downloaded from the miRBase (http://www.mirbase.org/) (release 17). We searched an EMBL format file of miRBase by using an "experimental" keyword to select only experimentally validated animal miRNAs. Finally the redundant miRNA sequences were removed. The remaining miRNAs were used as the reference set. The ESTs were obtained from the *Penaeus monodon* EST database (Tassanakajon *et al.,* 2006). We constructed the database using cDNA libraries obtained from eyestalk, hepatopancreas, haematopoietic, haemocyte, lymphoid organ, and ovary tissues of the shrimp. This resulted in a total of 10,536 unique sequences with 3,227 overlapping contigs and 7,309 singletons.

2.2 Computational identification of *Penaeus monodon* miRNA genes

The workflow for the computational identification in *Penaeus monodon* is shown in Figure 1. Both strands of the *Penaeus monodon* ESTs were used as query sequences for BLAST searches against sequences of experimental validated animal miRNAs. We used default parameters for the BLAST searches, except for the E-value and word-match size which were set to 10 and 7, respectively (Wang *et al.,* 2005). Only EST sequences with less than three mismatches to a mature miRNA were used. For the evaluation of hairpin structures, we extracted 100 nt upstream and 100 nt downstream sequences of each BLAST hits. In the case that the total length of the sequence was shorter than 200 nt, we used the entire sequence of a putative miRNA precursor sequence. All simple repeated regions were removed by using the RepeatMasker algorithm (Smit *et al.*, 1996) because such regions may fold into structures similar to pre-miRNA hairpins. The coding regions were removed from these putative

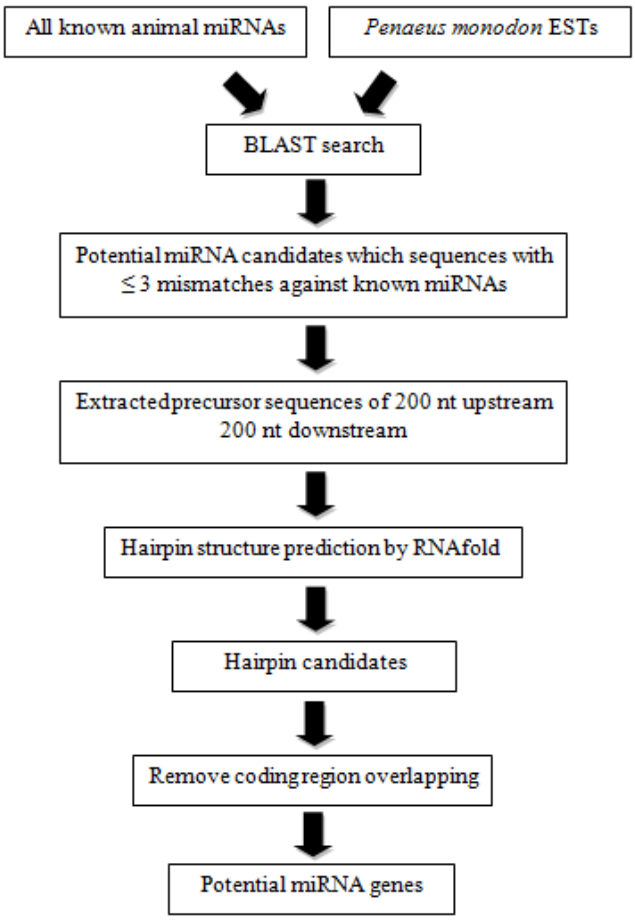

Figure 1. Workflow for computational identification of potential miRNA gene sequences.

Table 1. New miRNAs identified in *Penaeus monodon*.

miRNA	EST (Strand)	NM	LM	LP	MFE	miRNA sequence (5' to 3')
miR-4152	CT1439 (-)	2	18	100	-22.20	AGAUGUAGUUACUGUAAA
miR-466	CT1441 (-)	3	23	71	-20.70	UGUGUGUAUACAUGUAUAAGUGA
kmiR-32*	SG3984 (-)	3	21	125	-33.40	CAAUUUAGUGUGUGUGUGAUU
lin-4	SG4975 (-)	3	20	91	-23.70	UCCCUGAGACUUCAACUCUG
miR-1346	SG8776 (-)	1	18	104	-26.80	GUGGGUUGGGGGGGGGG
miR-4310	SG9237 (+)	1	16	96	-33.42	GCAGCAUUGAUGUCCC

NM: Number of mismatches; LM: Length of mature miRNA; LP: Length of precursor; MFE: Minimal folding free energy (kcal/mol).

precursor sequences by searching against a NCBI non-redundant (nr) protein database using BLASTX (Altschul *et al.,* 1997). The putative precursor sequences which had no hits at an E-value of 10^{-6} and all overlaps with any known proteins, were treated as non-coding ESTs and kept for the next step of the structural analysis. The secondary structures of putative pre-miRNAs were predicted by RNAfold in the Vienna RNA package (Hofacker, 2004; Hofacker, 2009). Only the sequences which satisfied the following criteria (Ambros *et al.,* 2003; Wang *et al.,* 2005) were treated as conserved miRNAs:

(1) The free energy of predicted secondary structures was less than -20 kcal/mol,

(2) Putative miRNAs were located on the same arm of the precursor as the matched homolog,

(3) A minimum of 16 base pairs existed between miRNAs and the matched miRNA*.

3. Results and Discussion

3.1 Potential *Penaeus monodon* miRNAs

We can identify six miRNA in *Penaeus monodon* by using the workflow of homology searches and structural analysis as shown in Table 1. The length of the predicted miRNAs was in the range from 16 nt to 23 nt, while the length of the predicted precursor sequences was fluctuated ranging from 71 nt to 125 nt with an average of 98 nt. These sequences folded into a typical stem-loop structure, having the mature miRNA on the 5′, or alternatively on the 3′ end of an EST (Figure 2). The hairpin loop secondary structures had a minimum folding free energy ranging from -33.42 kcal/mol to -22.20 kcal/mol.

The mir-4152 precursor, mir-466k precursor, mir-32v precursor, lin-4 precursor, mir-1346 precursor and mir-4310 precursor were encoded as miR-4152-5p (10 nt to 17 nt), miR-466k (1 nt to 23 nt), miR-32* (52 nt to 82 nt), lin-4 (5 nt to 24 nt), miR-1346 (86 nt to 103 nt) and miR-4310 (71 nt to 86 nt), respectively. miR-1346 and miR-4310 were the most conserved containing only one mismatch with their homologues. All mismatch positions occurring in the predicted miRNAs were located outside the seed region (2 nt to 7 nt). The seed regions usually perfectly match with the targeted mRNAs. This suggests that the miRNAs tend to target with the same mRNA and have the same function to their homologues. In previous work on human data, about 14000 ESTs revealed one miRNA (Li *et al.,* 2006) while in this work, 6 miRNAs was predicted from the total of 10,536 unique EST. These miRNAs might be a result of the method which was used to build up the ESTs or they might be transcripted together as a polycistronic microRNA cluster.

Among the six mature miRNAs, mir-4152, mir-466k, miR-32*, lin-4, mir-1346 and mir-4310, only lin-4 had been previously studied experimentally. Lin-4 was recognized as the first miRNA, discovered during a study of the gene lin-14 in the development of *Caenorhabditis elegans* by Ambros

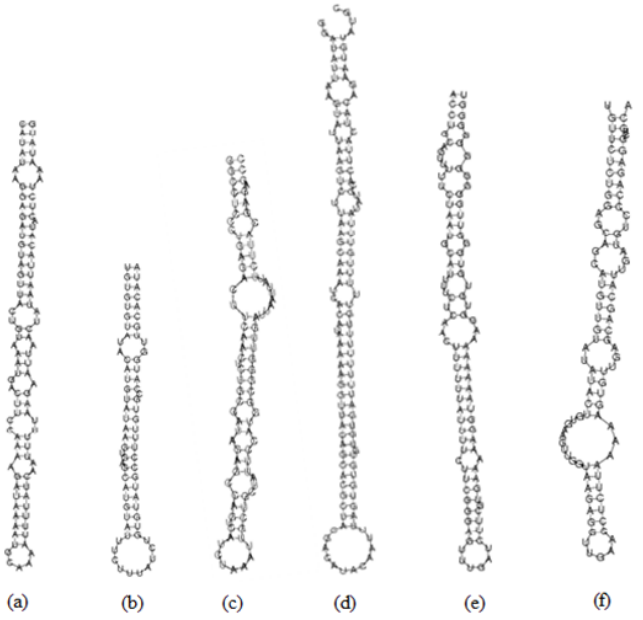

Figure 2. Predicted hairpin loop structures of newly identified *Penaeus monodon* miRNA precursors. (a) mir-4152 precursor (miR-4152-5p (10 to 17)). (b) mir-466k precursor (miR-466k (1 to 23)). (c) mir-32 precursor (miR-32* (52 to 82)). (d) lin-4 precursor (lin-4 (5 to 24)). (e) mir-1346 precursor (miR-1346 (86 to 103)). (f) mir-4310 precursor (miR-4310 (71 to 86)). The position of each miRNA is indicated by the number in parentheses.

and his colleague (Lee *et al.,* 1993). They found that lin-14 protein was down-regulated by a 22 nt RNA, lin-4, that contained sequences partially complementary to multiple sites in the 3' UTR of the lin-14 mRNA. Another study indicated that lin-4 may play an important role in regulating fat accumulation and control of the life span of *Caenorhabditis elegans* via production of reactive oxygen species (ROS) (Zhu *et al.,* 2010).

3.2 Prediction of the miRNA targets

In the preliminary study of the miRNAs targets in *Penaeus monodon* using the miRanda algorithm (Enright *et al.,* 2003) a number of potential targets for mir-466k, lin-4, mir-1346 and mir-4310 were identified (Table 2). In total, 25 target sites for 22 potential target genes were identified. There were 2 predicted target genes in *Penaeus monodon* which had more than one predicted target sites for a given miRNA. SG5444 (PREDICTED: T-cell receptor beta chain ANA 11) and SG5916 (PREDICTED: Histidine rich glycoprotein precursor) were hit by the miR-466k at three and two sites, respectively.

mir-466k had the potential to target a gene of the juvenile hormone esterase which is related to the regulation of molting, metamorphosis, reproductive maturation, and pheromone biosynthesis in insects (Tsubota *et al.,* 2010). Moreover the mir-466k had the potential to target genes belonging to the GTPase superfamily. GTPase was associated with both constitutive and regulated secretory pathways, and might be involved in protein transport. Understanding the mir-466k regulatory pathway may lead to applications for artificial seed production and enhancement of growth in shrimp aquaculture.

In *Penaeus monodon*, lin-4 probably targeted Hemocyanin, the main protein component of the hemolymph, which was related to the innate immune response of the crustacean (Cheng *et al.* 2008). Furthermore lin-4 was a potential target for GTP-binding protein involving in a conformational change mediated by the hydrolysis of GTP to GDP (Moller *et al.,* 1987). This family of proteins promoted the GTP-dependent binding of aminoacyl tRNA to the A site of ribosomes during protein biosynthesis. The protein also catalyzed the translocation of the synthesized protein chain from the A to the P site (Stansfield *et al.,* 1995).

There were two interesting putative targeted genes for miR-1346. The first one, NSFL1 cofactor p47, was required for the membrane reassembly of the endoplasmic reticulum, the nuclear envelope and the Golgi apparatus at the end of mitosis. The other one was a C-type lectin which played a key role in pathogen recognition, innate immunity, and cell-cell

Table 2. The potential targets of newly identified miRNAs in *Penaeus monodon*.

miRNA	Targeted EST	Targeted Gene (target site no.)	Possible function	
mir-4152	-	-	-	-
mir-466	CT224	juvenile hormone esterase (1)	DNA replication	3UTR
	CT605	RAB family GTPase (1)	Endocytosis	3UTR
	CT2632	hypothetical protein (1)	carbohydrate metabolism	ORF
	SG4153	hypothetical protein (1)	DNA repair	5UTR
	SG4281	hypothetical protein (1)	DNA replication, DNA repair	ORF
	SG4325	DEAH-box RNA helicase (1)	lipase activity	ORF
	SG4793	T-cell receptor beta chain ANA 11 (1)	DNA replication, DNA repair	ORF
	SG5444	T-cell receptor beta chain ANA 11 (3)	lipase activity	ORF
	SG5916	Histidine-rich glycoprotein precursor (2)	peptidoglycan biosynthesis	ORF
	SG8549	mCG1041337 (1)	lipase activity	3UTR
kmiR-32*	-	-	-	-
lin-4	CT219	Hemocyanin (1)	methionyl-tRNAaminoacylation	ORF
	SG4820	hypothetical protein (1)	motor activity, transport	ORF
	SG8252	hypothetical protein (1)	ciliary or flagellar motility	ORF
	SG10292	GTP-binding protein (1)	oxidoreductase activity	ORF
miR-1346	CT137	Tubulin alpha-3 chain (1)	phosphorelay	5UTR
	CT161	Transglutaminase (1)	RNA processing	3UTR
	CT568	C-type lectin 3 (1)	thiamin biosynthesis	ORF
	CT1474	regulator of g protein signaling (1)	phosphorelay	3UTR
	CT3028	alpha-endosulfine (1)	DNA replication	ORF
	SG6162	annexin (1)	Metabolism	ORF
	SG6206	hypothetical protein (1)	polysaccharide biosynthesis	ORF
mir-4310	CT27	tubulin alpha chain (1)	nitrate transport	3UTR

interactions (Robinson *et al.,* 2006). The protein might be capable of activating or inhibiting receptors involving in the innate responses to pathogens.

4. Conclusions

The *Penaeus monodon* genome encoded at least six miRNA orthologs that were common to 14 other species. Further experiments are now in progress to verify the expression of the predicted miRNAs and to evaluate their target genes in *Penaeus monodon*. We have developed a computational workflow for automatically identifying conserved miRNAs by using the sequence and structural homology search strategy. This can be used for consecutive identification, as more sequences become available. The Perl script of the workflow is freely available on request. Therefore, the workflow package and the findings from this study would be useful for other research concerned with the function of *Penaeus monodon* miRNAs and their regulatory mechanisms.

Acknowledgements

This work was supported by the Grants from Prince of Songkla University. The authors also thank Professor Anchalee Tassanakajon, and National Center for Genetic Engineering and Biotechnology (BIOTEC), for the *Penaeus monodon* EST data.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller W. and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 25(17), 3389-3402.

Ambros, V. 2004. The functions of animal microRNAs. Nature. 431(7006), 350-355.

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., Dreyfuss, G., Eddy, S. R., Griffiths-Jones, S., Marshall, M., Matzke, M., Ruvkun, G. and Tuschl, T. 2003. A uniform system for microRNA annotation. RNA. 9(3), 277-279.

Bartel, D. P. 2004. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 116(2), 281-297.

Cheng, W., Tsai, I. H., Huang, C. J., Chiang, P. C., Cheng, C. H. and Yeh, M. S. 2008. Cloning and characterization of hemolymph clottable proteins of kuruma prawn (*Marsupenaeus japonicus*) and white shrimp (*Litopenaeus vannamei*). Developmental & Comparative Immunology. 32(3), 265-274.

Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D. S. 2003. MicroRNA targets in Drosophila. Genome Biology. 5(1), R1.

Griffiths-Jones, S. 2004. The microRNA Registry. Nucleic Acids Research. 32 (Database issue), D109-111.

Griffiths-Jones, S. 2006. miRBase: the microRNA sequence database. Methods in Molecular Biology. 342, 129-138.

Griffiths-Jones, S. 2010. miRBase: microRNA sequences and annotation. Current Protocols in Bioinformatics. Chapter 12, Unit 12 19 11-10.

Griffiths-Jones, S., Grocock, R. J., S. van Dongen, Bateman, A. and Enright, A. J. 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Research. 34 (Database issue), D140-144.

Griffiths-Jones, S., Saini, H. K., S. van Dongen and Enright, A. J. 2008. miRBase: tools for microRNA genomics. Nucleic Acids Research 36. (Database issue), D154-158.

Hofacker, I. L. 2004. RNA secondary structure analysis using the Vienna RNA package. Current Protocols in Bioinformatics. Chapter 12, Unit 12 12.

Hofacker, I. L. 2009. RNA secondary structure analysis using the Vienna RNA package. Current Protocols in Bioinformatics. Chapter 12, Unit12 12.

Huang, T., Xu, D., Zhang, X. 2012. Characterization of host microRNAs that respond to DNA virus infection in a crustacean. BMC Genomics. 13, 159.

Huang, Y., Shen, X. J., Zou, Q., Wang, S. P., Tang, S. M. and Zhang, G. Z. 2011. Biological functions of microRNAs: a review. Journal of Physiology and Biochemistry. 67 (1), 129-139.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W. and Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. Science. 294(5543), 853-858.

Lai, E. C. 2003. microRNAs: runts of the genome assert themselves. Current Biology. 13(23), R925-936.

Lee, R. C., Feinbaum, R. L. and Ambros, V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 75(5), 843-854.

Li, S. C., Pan, C. Y. and Lin, W. C. 2006. Bioinformatic discovery of microRNA precursors from human ESTs and introns. BMC Genomics. 7, 164.

Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B. and Bartel, D. P. 2003. Vertebrate microRNA genes. Science. 299 (5612), 1540.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S. Rhoades, M. W., Burge, C. B. and Bartel, D. P. 2003. The microRNAs of Caenorhabditis elegans. Genes & Development. 17(8), 991-1008.

Moller, W., Schipper, A. and Amons, R. 1987. A conserved amino acid sequence around Arg-68 of Artemia elongation factor 1 alpha is involved in the binding of guanine nucleotides and aminoacyl transfer RNAs. Biochimie. 69(9), 983-989.

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B. and Bartel, D. P. 2002. MicroRNAs in plants. Genes & Development. 16(13), 1616-1626.

Robinson, M. J., Sancho, D., Slack, E. C., LeibundGut-Landmann, S. and Reis e Sousa, C. 2006. Myeloid C-

type lectins in innate immunity. Nature Immunology. 7(12), 1258-1265.

Ruan, L., Bian, X., Ji ,Y., Li, M., Li, F. and Yan, X. 2011. Isolation and identification of novel microRNAs from *Marsupenaeus japonicus*. Fish Shellfish Immunol. 31, 334-340.

Smit, A. F. A. , Hubley, R. and Green, P. RepeatMasker Open-3.0. 1996-2010. (http://www.repeatmasker.org)

Stansfield, I., Jones, K. M., Kushnirov, V. V., Dagkesamanskaya, A. R., Poznyakovski, A. I., Paushkin, S. V., Nierras, C. R. Cox, B. S., Ter-Avanesyan, M. D. and Tuite, M. F. 1995. The products of the SUP45 (eRF1) and SUP35 genes interact to mediate translation termination in *Saccharomyces cerevisiae*. EMBO Journal. 14(17), 4365-4373.

Tassanakajon, A., Klinbunga, S., Paunglarp, N. Rimphanitchayakit, V., Udomkit, A., Jitrapakdee, S., Sritunyalucksana, K., Phongdara, A., Pongsomboon, S., Supungul, P., Tang, S., Kuphanumart, K., Pichyangkura, R. and Lursinsap, C. 2006. Penaeus monodon gene discovery project: the generation of an EST collection and establishment of a database. Gene. 384, 104-112.

Tsubota, T., Minakuchi, C., Nakakura, T., Shinoda, T. and Shiotsuki, T. 2010. Molecular characterization of a gene encoding juvenile hormone esterase in the red flour beetle, *Tribolium castaneum*. Insect Molecular Biology. 19(4), 527-535.

Wang, X., Zhang, J., Li, F., Gu, J., He, T., Zhang, X. and Li, Y. 2005. MicroRNA identification based on sequence and structure alignment. Bioinformatics. 21(18), 3610-3614.

Weber, M. J. 2005. New human and mouse microRNA genes found by homology search. FEBS Journal. 272(1), 59-73.

Wightman, B., Ha, I. and Ruvkun, G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. Cell. 75(5), 855-862.

Williams, A. E. 2008. Functional aspects of animal microRNAs. Cellular and Molecular Life Sciences. 65(4), 545-562.

Zhu, C., Ji, C. B., Zhang, C. M., Gao, C. L., Zhu, J. G., Qin, D. N., Kou, C. Z., Zhu, G. Z., Shi, C. M. and Guo, X. R. 2010. The lin-4 Gene Controls Fat Accumulation and Longevity in Caenorhabditis elegans. International Journal of Molecular Sciences .11(12), 4814-4825.