Computational classification of small RNAs and their targets

Simon Moxon

Supervisor: Prof. Vincent Moulton
Co-supervisor: Dr. Tamas Dalmay

A thesis submitted for the Degree of Doctor of Philosophy at the University of East Anglia

August 2008

School of Computing Sciences
University of East Anglia
Norwich, United Kingdom

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification at this or any other university or other institute of learning.

Acknowledgements

Firstly I would like to thank my supervisor, Vincent Moulton, and my co-supervisor Tamas Dalmay for the valuable advice and support that they gave me during my PhD. I would also like to thank Frank Schwach for sharing his office and knowledge with me.

I would like to thank David Studholme and David Baulcombe for giving me the opportunity to spend some time over at the Sainsbury Laboratory and to learn from them, and also to acknowledge Alex Bateman and Sam Griffiths-Jones for the encouragement and confidence they gave me to pursue a PhD in the first place. I would like to thank the School of Computing Sciences at UEA for their financial support through the scholarship they provided, and for giving me the opportunity to undertake this work.

Finally I would like to thank my wife Manuela and my parents, Dee and John, for their support and understanding over the last three years.

Publications

Pilcher, R., **Moxon, S.**, Pakseresht, N., Moulton, V., Manning, K., Seymour, G., Dalmay, T. (2007): Identification of novel small RNAs in tomato (*Solanum lycopersicum*). *Planta* **226**(3):709-17.

Moxon, S., Moulton, V., Kim, J.T. (2008): A scoring matrix approach to detecting miRNA target sites. *Algorithms for Molecular Biology*. **3**(1):3.

Bagnall, A., **Moxon, S.**, Studholme, D. (2008): Time Series Data Mining Algorithms for Identifying Short RNA in *Arabidopsis thaliana*. *BIOCOMP'08 - The 2008 International Conference on Bioinformatics & Computational Biology*. In press

Moxon, S.[*], Jing, R.[*], Szittya, G.[*], Schwach, F., Pilcher, R. L. R., Moulton, V., Dalmay, T. (2008): Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. *Genome Research*. **18**(10):1602-9.

Moxon, S.[*], Schwach, F.[*], Studholme, D., Dalmay, T., MacLean, D., Moulton, V. (2008): A toolkit for analysing large-scale plant small RNA datasets *Bioinformatics*. **24**(19):2252-3.

Szittya, G.[*], **Moxon, S.**[*], Santos, D. M., Jing, R., Fevereiro, M. P. S., Moulton, V., Dalmay, T. (2008): *BMC Genomics*. In press

[*] Authors contributed equally to this work.

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, and that any ideas or quotations from the work of other people, published or otherwise, are fully acknowledged.

Abstract

Small RNAs, and in particular microRNAs, are currently receiving a great deal of attention due to their important roles in gene regulation and organism development. Recently, new high-throughput technologies have made it possible to sequence hundreds of thousands of small RNAs from a single experimental sample. In this thesis we develop new computational tools to process such high-throughput small RNA datasets in order to identify microRNAs and other biologically interesting small RNA candidates and to predict their target genes. We apply these tools to a variety of plant and animal datasets and present some novel discoveries including miRNAs involved in fruit development in tomato (*Solanum lycopersicon*).

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The focus of the research presented in this thesis is the development of new computational tools and algorithms to enable the detection and classification of novel, biologically interesting, small RNAs from experimental sequence data. In addition, applications of these tools are presented, and a number of interesting biological discoveries identified through computational predictions and analyses are reported. Much of the work presented in this thesis has involved collaborations with biologists who have provided both small RNA sequence data and experimental validation of computational predictions. An overview of the thesis is given below along with details of collaborations.

Chapter 2. We provide background information about small RNAs and explain their function and biogenesis. We then go on to describe both current computational and laboratory based methods for small RNA detection and classification which are important in later chapters.

Chapter 3. We implement a tool, `miRCat`, to classify miRNA sequences in both animal and plant high-throughput small RNA datasets. At the time of implementation it was the only tool specifically designed for this purpose, although, since then another very similar algorithm `miRDeep` [Friedländer et al., 2008] has been published that is only applicable to animal datasets. We then go on to describe the first online resource for researchers to upload and process their high-throughput small RNA datasets: `http://srna-tools.cmp.uea.ac.uk`. The `SiLoCo` and RNAfolding/annotation tools detailed in this chapter were created by Dr Frank Schwach and a description of these tools is included in the thesis for completeness.

Chapter 4. We use computational methods to identify a number of putative miRNA sequences and to find miRNA homologues conserved between *Arabidopsis thaliana* and tomato (*Solanum lycopersicon*). This work was carried out whilst developing the `miRCat` tool described in Chapter 3, and discoveries made during this work enabled the refinement of the software. All laboratory-based experimental work in this chapter was carried out by members of Dr. Tamas Dalmay's group.

Chapter 5. We apply methods described in Chapter 3 to analyse high-throughput small RNA sequence sets in *Solanum lycopersicon*. Here we identify the first tomato specific miRNAs and their targets as well as miRNAs believed to be involved in the control of tomato fruit ripening. We also describe both an interesting class of transposon derived small RNAs and a set of small RNAs derived from a virus that is integrated into the host genome. Both observations were made using bioinformatic,

sequence analysis techniques. All experimental work in this chapter was carried out by Dr. Tamas Dalmay's group.

Chapter 6. We describe a novel computational approach for the identification of miRNAs from high-throughput small RNA sequence sets. This method does not rely on a genome sequence being available. Instead it classifies putative miRNA/miRNA* pairs from an input dataset using a Support Vector Machine approach. This technique is the first miRNA prediction algorithm that does not rely on a genome sequence, making it an invaluable tool when working with organisms where published sequence data is not available. All laboratory work in this chapter was carried out by Dr. Jing Runchun.

Chapter 7. We introduce a novel miRNA target prediction algorithm `StackBM` which is the first method that is applicable to both animal and plant data. It makes use of previously experimentally validated target sites in the search for novel targets. We show that it performs well in terms of sensitivity and specificity in comparison with a previously published target prediction method, and discuss future extensions of this work. Work in this chapter was based on the Binding Matrix technique for transcription factor binding site classification [Kim et al., 2004] and the `StackBM` software was implemented by Dr. Jan Kim. The idea for using multiple matrices to incorporate gaps in input alignments, all experimental testing, refinements of the method and generation of results were my contribution to this work.

Chapter 8. We discuss work presented in this thesis and go on to detail possible future directions and extensions to the research.

Chapter 2

Background

2.1 Summary

This chapter provides a brief introduction to small RNAs, detailing both their biogenesis, function and importance in biological systems. It then goes on to describe traditional laboratory-based techniques to sequence, analyse and validate small RNAs, including an overview of the new high-throughput sequencing technologies that are currently emerging. Current state of the art computational methods for small RNA detection, classification and target prediction are then discussed, and an outline of the strengths and weaknesses of the various approaches is provided, this gives the necessary background to the work presented in the rest of the thesis.

2.2 Small RNAs

Small RNA (sRNA) is a general term applied to a broad class of short non-coding RNAs (ncRNAs) that are not translated into a protein product but instead function directly at the level of the RNA in the cell. sRNAs are typically between 18 and 30

nucleotides in length and are involved in gene regulation and genome maintenance. sRNAs are derived from either double-stranded RNA (dsRNA) or highly structured precursor sequences and be subdivided into several distinct groups based on their properties, functions and biogenesis. See [Kim, 2005b], [Chapman and Carrington, 2007], and [Kawaji and Hayashizaki, 2008] for recent overviews.

Several important classes of sRNA are described in detail below.

2.2.1 MicroRNAs

MicroRNAs (miRNAs) are a class of sRNA, typically between 21 and 24 nucleotides in length [Lau et al., 2001]. The first miRNA to be discovered was *lin-4*, found in the nematode worm *Caenorhabditis elegans* [Lee et al., 1993]. *lin-4* was thought to be a biological peculiarity [Pasquinelli, 2002] and was for many years the only example of such a sRNA. In 2000 Reinhart *et al*. discovered a second *C. elegans* miRNA *let-7* [Reinhart et al., 2000], involved in developmental timing, which was found to be conserved in many other animals including humans and the fruit fly *Drosophila melanogaster* [Pasquinelli et al., 2000]. In 2001 several groups identified a number of new *C. elegans* miRNAs [Lau et al., 2001, Lee and Ambros, 2001, Lagos-Quintana et al., 2001] which firmly established that miRNAs were important regulatory molecules and represented a whole new level of gene regulation that had previously been overlooked.

In 2002 the first plant miRNAs were identified in *Arabidopsis thaliana*

[Reinhart et al., 2002] and in 2004 miRNAs were identified in the Epstein-Barr virus (EBV) [Pfeffer et al., 2004]. Since then many new miRNAs have been discovered in a variety of animals [Lagos-Quintana et al., 2002, Lagos-Quintana et al., 2003, Xu et al., 2003, Wienholds et al., 2005], plants [Wang et al., 2004a, Lu et al., 2005b, Jin et al., 2008, Subramanian et al., 2008] and viruses [Grundhoff et al., 2006, Grey et al., 2005, Cai et al., 2005], but not in fungi or bacteria which are thought to lack the machinery required for miRNA production.

Functional miRNAs are derived from a longer, single stranded, primary RNA transcript known as a pri-miRNA which can be several kilobases in length [Bartel, 2004, Kim, 2005a]. In animals the pri-miRNA is processed by the Microprocessor complex to form a short (around 70nt) sequence which is able to fold into an imperfect stem-loop structure known as the pre-miRNA [Gregory et al., 2004]. In plants the pri-miRNA is processed by the Dicer-like 1 (DCL1) enzyme to yield a pre-miRNA hairpin. Plant pre-miRNAs also undergo extensive base-pairing to form a hairpin-like structure (see Figure 2.1), and are generally longer than their animal counterparts [Reinhart et al., 2002].

The pre-miRNA is further processed by DCL1, HEN1 and HYL1 enzymes [Schauer et al., 2002, Park et al., 2002, Han et al., 2004] in plants and Dicer in conjunction with a specialised RNA-binding protein e.g. Loquacious (Loqs) in flies or trans-activation responsive RNA-binding protein (TRBP) in humans

[Förstemann et al., 2005]. The miRNA along with its complement (known as the miRNA*) are excised from the hairpin where they form a double stranded RNA (dsRNA) intermediate. The mature miRNA sequence is the bound to an Argonaute protein (AGO1 in *Arabidopsis*) [Vaucheret et al., 2004] and recruited into the RNA-induced silencing complex (RISC) where it acts to regulate genes by binding to sites complementary to the miRNA sequence on messenger RNAs (mRNAs) [Reinhart et al., 2000], while the miRNA* strand gets degraded or accumulates at a lower level [Jones-Rhoades et al., 2006].

The translation of the mRNA is inhibited by the binding of the miRNA/RISC. Regulation is achieved either by translational repression (blocking the protein production machinery) or mRNA cleavage at a specific position in the miRNA/target duplex followed by degradation of the cleaved transcript. The method of regulation is dependent on the level of complementarity between miRNA and target site [Aukerman and Sakai, 2003], with highly complementary miRNA/target duplexes leading to mRNA cleavage and degradation (predominant in plants) and poor complementarity leading to translational repression (predominant in animals).

Mature miRNAs are often perfectly conserved between a wide variety of organisms since there is strong selection pressure for the conservation of the sequence. However, miRNA hairpins often differ significantly outside of the miRNA and miRNA* regions as there the structure rather than sequence must be conserved. This means that compensatory mutations that change the sequence but conserve the structure

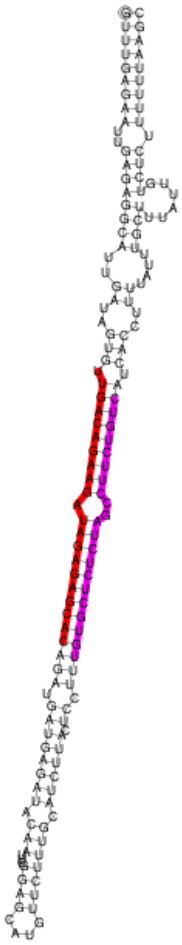

Figure 2.1: Secondary structure plot of the *Arabidopsis thaliana* miR157 pre-miRNA. The mature miRNA is shown in red and the miRNA* sequence is represented in pink.

are common.

miRNAs are known to be involved in stem cell differentiation, organ development, cell signaling, stress response and cancer [Zhang et al., 2007, Válóczi et al., 2006, Kloosterman and Plasterk, 2006]. This diverse range of roles coupled with their recent discovery has led to an increased scientific interest in their identification and analysis. New miRNAs are now characterised on a regular basis, with 6396 sequences

from 72 different species described in the latest release (11.0) of the central miRNA repository miRBase [Griffiths-Jones et al., 2008]. The exact number of miRNAs in any given organism is unknown although different figures have been proposed. Estimates in human range from around 1000 [Berezikov et al., 2005] to tens of thousands [Miranda et al., 2006]. 184 miRNAs have so far been discovered in the model plant *Arabidopsis thaliana* and recent studies have suggested that *Arabidopsis* miRNA genes undergo relatively frequent birth and death with only a small subset being evolutionarily conserved [Fahlgren et al., 2007]. This could mean that a large number of newly evolved, non-conserved miRNAs are yet to be classified.

Natural antisense miRNAs

Natural antisense miRNAs (nat-miRNAs) are a recently discovered class of plant miRNA [Lu et al., 2008] that undergo a distinct mechanism of biogenesis (see Figure 2.2). Overlapping sense and antisense transcripts are first produced from the nat-miRNA locus. The primary antisense transcript contains a large intron which must be spliced out in order to allow the transcript to fold into a typical pre-miRNA hairpin. The pre-miRNA then feeds into the regular miRNA processing pathway where the mature miRNA along with the miRNA* sequence are excised from the hairpin by DCL1. The mature miRNA goes on to target the sense transcript from the same locus causing mRNA cleavage, therefore regulating the expression of this gene.

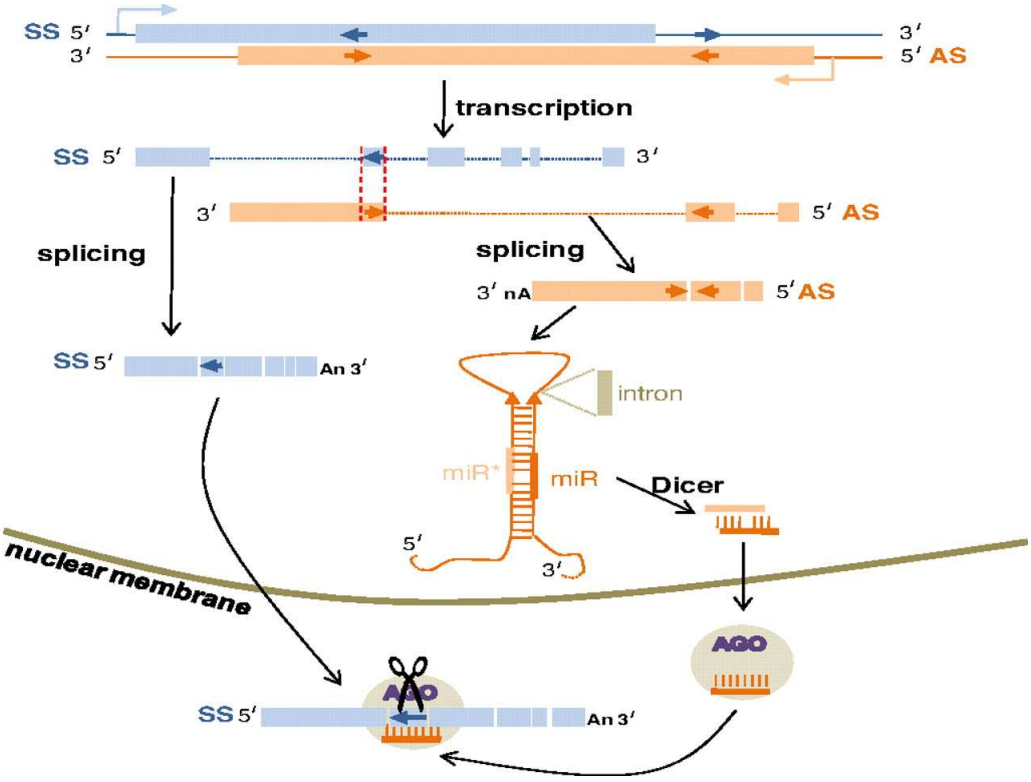

Figure 2.2: Model for nat-miRNA biosynthesis and function. The nat-miRNA pathway initiates with the splicing of pri-miRNA transcripts to yield pre-miRNA hairpins. It is believed that the possession of the introns in nat-miRNA precursors is important for the biosynthesis of miRNAs. Splicing of these introns limits the potential base-pairing of the pre-nat-miRNA with the sense transcript and favors hairpin formation. After Dicer cleavage, the mature nat-miRNAs then enter the cytoplasm and direct the cleavage of the sense transcripts that are their targets. AS, antisense strand; SS, sense strand. This figure is reproduced from [Lu et al., 2008].

2.2.2 Short interfering RNAs (siRNAs)

Unlike miRNAs, short interfering RNAs (siRNAs) are derived from long, perfectly complementary double-stranded RNAs (dsRNAs). siRNAs can be either endogenous (originating within an organism) or exogenous (originating outside of an organism). Examples of exogenous siRNAs are experimentally introduced dsRNA [Fire et al., 1998], transgenes [Voinnet et al., 1998] and viruses [Kasschau and Carrington, 1998], which can initiate an RNA silencing response in plants [Hamilton and Baulcombe, 1999] and some animals such as *Caenorhabditis elegans* whereby distinct siRNA populations are formed during primary and secondary phases. Primary siRNAs are produced from the initial exogenous trigger whereas secondary siRNA production can be initiated by primary siRNAs targeting a transcript and inducing the production of secondary siRNAs [Sijen et al., 2007, Mlotshwa et al., 2008, Moissiard et al., 2007] which can spread throughout a tissue or organism [Voinnet, 2005].

Several subtypes of endogenous siRNAs are currently known and have distinct regulatory functions described below.

Trans-acting short interfering RNAs

trans-acting siRNAs (ta-siRNAs) [Vazquez et al., 2004] are found in plants but not in animals and, like miRNAs, are generally around 21nt in length. ta-siRNAs are produced from specific genomic loci called TAS loci. TAS genes are transcribed into mRNA and are cleaved in two positions by miRNAs. After cleavage they are made

double stranded by RNA-dependent RNA polymerase 6 (RDR6) (in *Arabidopsis*) and the double stranded sequence is then processed by Dicer-like 4 (DCL4) which sequentially cuts the transcript yielding “phased” siRNA that are in a 21nt register relative to the miRNA cleavage sites [Axtell et al., 2006, Allen et al., 2005] (Figure 2.3). The processed 21nt ta-siRNAs are then, like miRNAs, incorporated into the RISC where they act to target mRNAs which are then subject to regulation.

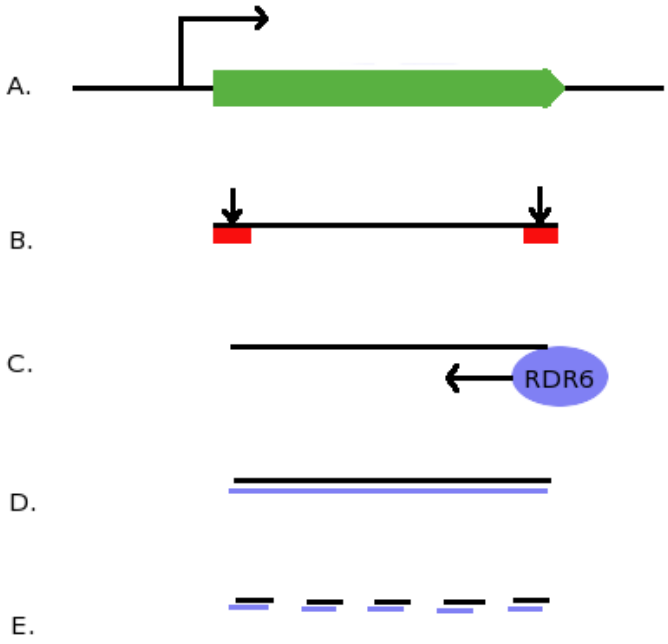

Figure 2.3: In-phase processing of ta-siRNA from TAS gene. A) TAS gene is transcribed giving rise to primary TAS transcript. B) The transcript is targeted and cleaved in two positions by miRNAs complementary to regions shown in red (indicated by arrows). C) and D) RDR6 recognises the processed transcript and makes the RNA double-stranded. E) DCL4 recognises the double stranded RNA and processes it into 21nt ta-siRNAs. Each ta-siRNA is in a 21nt phasing group relative to the initial miRNA cleavage position (ta-siRNAs are produced from positions 1-21, 22-43, 44-65 and so on).

Increasingly complex sRNA gene regulatory systems have been uncovered whereby ta-siRNAs can invoke a cascade of gene regulation. In the example described by Chen *et al*. [Chen et al., 2007] in *Arabidopsis*, miR173 cleaves the TAS2 transcript and causes the production of phased ta-siRNAs. One of the TAS2-derived ta-siRNAs, ta-siR2140, targets two pentatricopeptide repeat (PPR) genes, At1g63130 and At1g63080. The cleavage products of At1g63130 and At1g63080 then produce secondary ta-siRNAs, with the phase set by ta-siRNA rather than by miRNAs. One of the secondary ta-siRNAs, siR9as, is known to target another PPR gene, At1g62930, causing transcript cleavage leading to mRNA degradation and therefore regulating the gene.

Ta-siRNAs have been found to be an important regulator of plant development [Fahlgren et al., 2006, Nogueira et al., 2007, Peragine et al., 2004] and mutant plants lacking the necessary machinery to produce ta-siRNAs show accelerated juvenile to adult phase change in *Arabidopsis* which leads to severely deformed, stunted phenotypes.

Repeat associated short interfering RNAs (ra-siRNAs)

Repeat associated siRNAs or ra-siRNAs are produced from long dsRNA which map to repetitive sequence regions on the genome such as transposons, long inverted repeats and dispersed repetitive elements [Slotkin and Martienssen, 2007]. Dicer-like 3 (DCL3) and RNA-dependent RNA-polymerase 2 (RDR2) are responsible for

the generation of this class of siRNA which tend to be longer than ta-si and miRNAs (24nt). While DCL3 functions as the ribonuclease to process dsRNA precursors, RDR2 is thought to function as a polymerase to form dsRNA from a primary single-stranded RNA (ssRNA) transcript [Xie et al., 2004]. siRNAs produced from inverted repeats with near perfect self-complementarity are naturally double-stranded and therefore do not require RDR2 activity and are instead directly processed by DCL3 [Xie et al., 2004].

Repeat associated siRNAs appear to be randomly produced from the dsRNA precursors rather than following the phased-pattern seen with ta-siRNAs. ra-siRNAs can trigger epigenetic effects at target loci and are associated with RNA-directed DNA methylation and chromatin remodelling [Xie et al., 2004, Qi et al., 2006, Chan et al., 2004]. In *Arabidopsis*, the 24nt siRNAs associate with Argonaute 4 (AGO4), which in conjunction with DNA-dependent RNA polymerase IVb (PolIVb) [Mosher et al., 2008] guide DNA methylation. In plants, recent studies have shown that sRNA mediated DNA methylation is required in order to silence transposons and has also been implicated in the regulation of plant development [Lippman and Martienssen, 2004, Slotkin et al., 2005, Zilberman et al., 2007, Zhang et al., 2006c, Liu et al., 2004, Vaughn et al., 2007]. Loss of methylation results in a genome-wide transcriptional reactivation of transposons.

Natural-antisense transcript derived short interfering RNAs (nat-siRNAs)

nat-siRNAs have recently been described in *Arabidopsis thaliana* [Borsani et al., 2005]. Here two genes are regulated in response to salt stress: SRO5 gene expression is induced while P5CDH expression decreases. Transcription from opposing promoters yields SRO5 and P5CDH transcripts with a 760nt antisense overlap in their respective 3' regions. 24 and 21nt sRNAs matching this antisense overlap region were isolated from salt-stressed plants [Sunkar and Zhu, 2004, Borsani et al., 2005] thus showing a novel mechanism of sRNA production from overlapping transcripts.

Borsani *et al*. proposed a model whereby SRO5 and P5CDH transcripts anneal at the antisense, overlap region to form a dsRNA substrate for DCL2 which produces 24nt siRNAs. The 24nt nat-siRNAs then guide the additional phased, DCL1-dependent cleavages of the dsRNA into 21nt nat-siRNAs. The nat-siRNAs generated then target the P5CDH transcripts for degradation. The expression of SRO5 is induced by salt and is required to initiate siRNA formation.

A second example of a nat-siRNA locus was recently described by Katiyar-Agarwal *et al*. [Katiyar-Agarwal et al., 2006] in *Arabidopsis*. Here nat-siRNA production is specifically induced by the bacterial pathogen *Pseudomonas syringae* and confers disease resistance to the plant. Unlike in the Borsani *et al*. example, DCL1 was found to be required for the formation of the initial 24nt nat-siRNA rather than DCL2.

Natural antisense transcripts account for up to 7.4% of annotated transcription

units in the *Arabidopsis* genome but such loci do not exhibit an increased likelihood to give rise to small RNAs based on published MPSS sRNA data [Henz et al., 2007]. However, as both examples of *Arabidopsis* nat-siRNAs were discovered in plants that were exposed to different stress conditions it is possible that nat-siRNA production is not widespread under normal growth conditions and that siRNAs are only produced when subject to specific environmental stresses or other triggers.

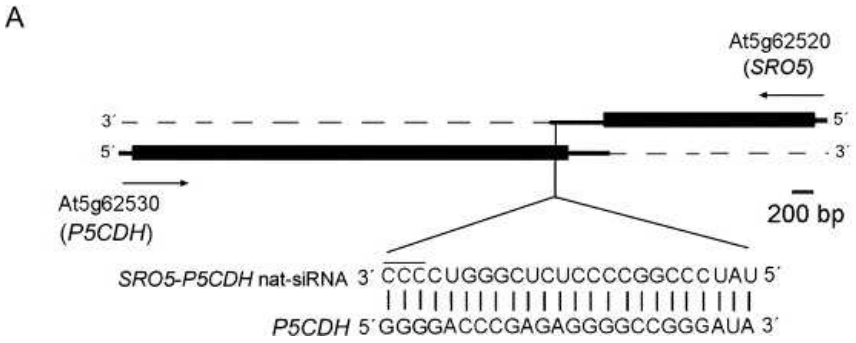

Figure 2.4: Salt stress induces a 24-nt nat-siRNA from the SRO5-P5CDH cis-antisense overlapping genes. Genomic structure of SRO5 (At5g62520) and P5CDH (At5g62530) genes. Arrows indicate the direction of transcription. Thick and thin solid lines represent ORF and UTR regions, respectively. Sequence of the 24-nt SRO5-P5CDH nat-siRNA is aligned with P5CDH mRNA. This figure is taken from [Borsani et al., 2005].

Piwi-interacting RNAs (piRNAs)

Piwi-interacting RNAs (piRNAs) are a class of small RNA molecules of between 29-30nt that are associated with a germline-specific subclass of Argonaute family proteins (Piwi proteins). Piwi/piRNA complexes are thought to be involved in transposon silencing in the germline genome of animals and are not present in plants.

The "ping-pong" model for piRNA production (Figure 2.5) has recently been proposed by two groups, (see Brennecke *et al*. [Brennecke et al., 2007] and Gunawardane *et al*. [Gunawardane et al., 2007] for details). This model shows how a piRNA cluster which gives rise to a variety of piRNAs can produce a sequence which is able to target a transposon thus causing cleavage and inactivating the transposon. This process also creates the 5' end of new AGO3-associated piRNA which is capable of cleaving complementary targets. One place from which such targets could be derived are the piRNA clusters themselves. Cleavage of cluster transcripts generates additional copies of the original antisense piRNA, which become available to silence active transposons. This process then becomes a self-amplifying loop where the initial response is general but, once a target is found, copies of the effective piRNA are amplified thus providing a specific response to the transposon.

2.3 Laboratory based methods for the detection and analysis of sRNAs

This section introduces the background to the laboratory-based analysis of sRNAs used later in this thesis.

2.3.1 sRNA detection by Northern blot analysis

The Northern blot is a technique used in molecular biology to detect RNA molecules with sequence specific probes, and it is often used for semi-quantitative assays where levels of RNA expression are compared using the intensity of bands produced on the

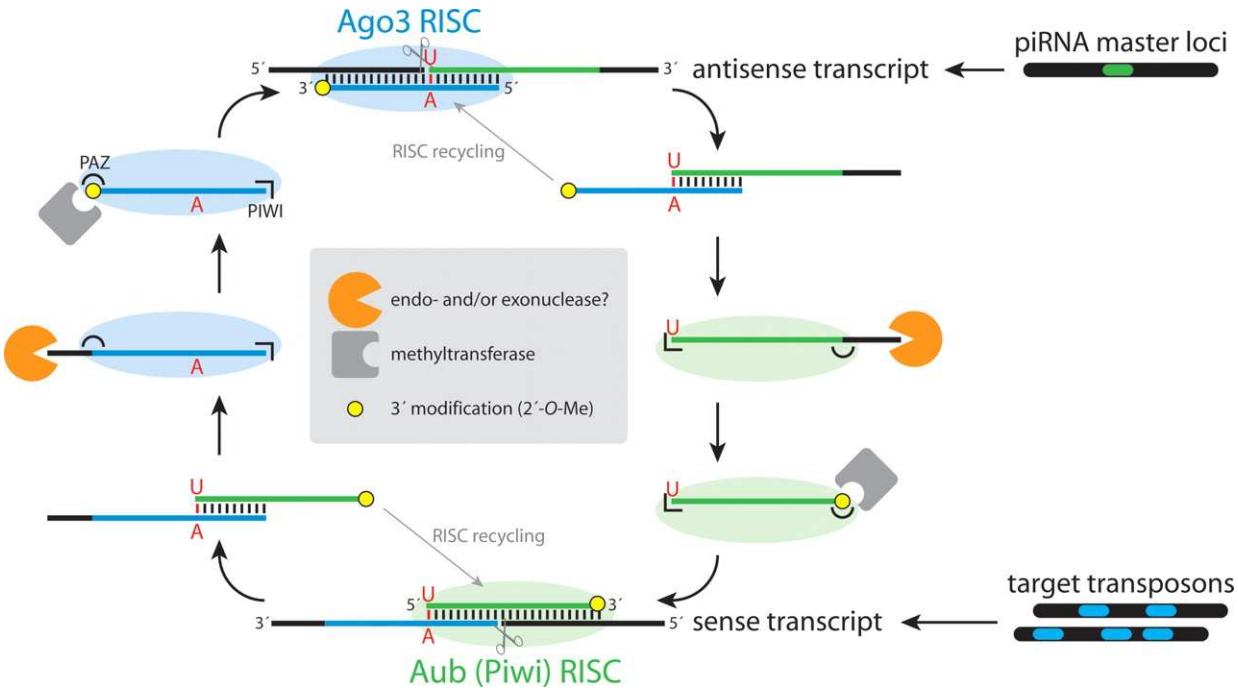

Figure 2.5: The ping-pong model for piRNA biogenesis. (Bottom) Sense transcripts from transposons are cleaved by Piwi or Aub RISC loaded with a piRNA guide. The cleaved transcript is not merely degraded but used to program Ago3 RISC. (Top) This complex in turn cleaves the antisense transcripts that originate from the master control loci. Again, the cleaved RNA serves to program Piwi or Aub RISC. Thus, sense and antisense transcripts fuel an amplification cycle in which the 5'-ends of piRNAs are defined by RISC cleavage. Presumably, the 3'-ends are shortened by an endonuclease and/or exonuclease to the size that fits the distance between PAZ and Piwi domains. The 3'-end is subsequently 2'-O-Me-modified by a methyltransferase, called Pimet/DmHen1 in *Drosophila*. This figure is taken from [Hartig et al., 2007].

blot.

Firstly RNA is purified from a biological sample. Next a probe, usually an oligonucleotide or an *in-vitro* transcript, is synthesised against a particular sequence of interest (e.g. a potential novel miRNA). The probe is complementary to the sequence of interest and is radioactively or chemiluminescently labeled so that it can be detected at a later stage.

The RNA sample is then loaded onto a gel for electrophoresis where an electric current is passed through the gel and the RNA moves away from the negative electrode. The distance moved depends on the size of the RNA fragment; shorter sequences move further down the gel, longer heavier sequences less so. This leads to the RNA molecules being separated according to their size and leaves a continuous smear on the gel. The gel is then transferred (blotted) onto a nitrocellulose membrane.

The probe is then added to the membrane and will hybridise to the single stranded sequence of interest if it is present in the RNA sample. The gel is washed to remove any non-specifically bound probe and then undergoes exposure. If the RNA of interest is present in the sample, then a dark band should be visible on the Northern blot, the intensity of which gives an estimate of the relative abundance of the molecule in the sample. An example Northern blot is shown in (Figure 2.6). Here several different samples were run on the same gel and the RNA of interest shows differing intensities or expression levels in the different developmental stages (strong expression in leaf and fruit 1-3mm, weaker in the other tissue samples).

2.3.2 Sequencing sRNAs

Most plant miRNAs have been identified by size fractionating (gel purification), ligating sRNAs (after adaptor ligation) into cloning vectors and then sequencing using the traditional Sanger sequencing method. This process was adopted in *Arabidopsis*, rice and poplar, and comparison of miRNA sequences across plant families showed

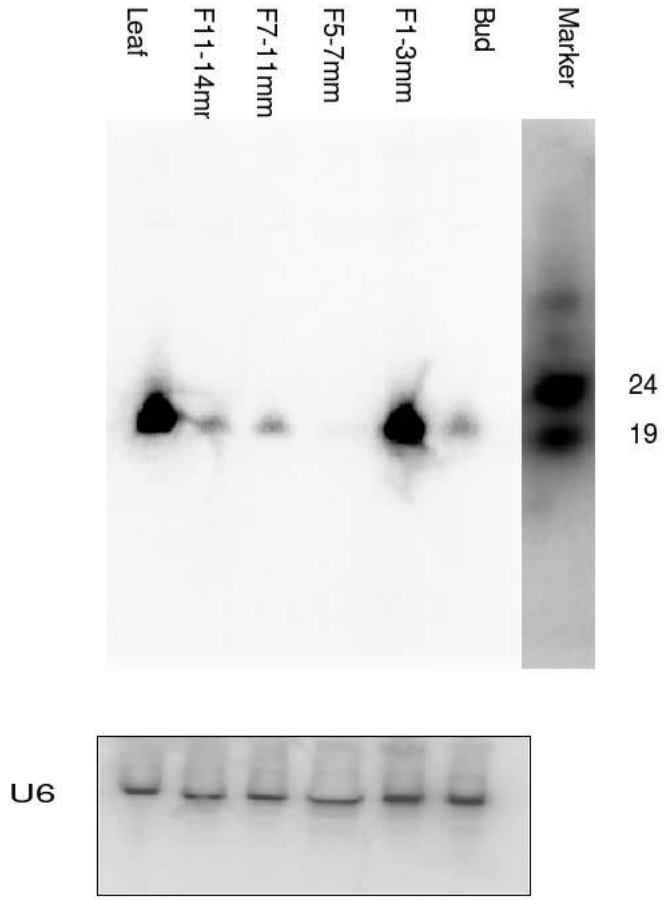

Figure 2.6: Example Northern blot of a candidate miRNA sequence in different plant tissues (leaf, fruit 11-14mm, fruit 7-11mm, fruit 5-7mm, fruit 1-3mm, bud).

that the majority were conserved [Axtell and Bartel, 2005]. Recently several new technologies have become available which have enabled high-throughput sequencing of sRNAs and have led to the discovery of a plethora of new miRNAs many of which are expressed at low levels and are either unique to a specific species or at least not widely conserved in related organisms.

Massively parallel signature sequencing (MPSS) [Reinartz et al., 2002] was the

first deep-sequencing method successfully used to discover a number of novel miRNAs in *Arabidopsis* [Wang et al., 2004b]. The MPSS sequencing technology identifies unique 17 nt sequences present in cDNA molecules originated from RNA extracted from a cell sample [Wang et al., 2004b, Brenner et al., 2000] the exact mechanisms of this process are not described in detail as MPSS data are not analysed in this thesis. The short read length of the MPSS system is not ideal for sRNA analysis as full length sRNAs cannot be sequenced making it harder to define individual sRNA species.

More recently, 454 pyrosequencing technology [Margulies et al., 2005] has been used by several groups [Lu et al., 2006, Fahlgren et al., 2007, Yao et al., 2007, Barakat et al., 2007a] in favour of MPSS. The 454 sequencing method generates up to around 400,000 reads per run and can generate much longer read lengths than MPSS (up to 300nt with the latest machines) meaning that complete sRNA reads can be sequenced.

In 454 sequencing specific 3' and 5' adaptors are added to the input sample and are necessary for purification, amplification, and sequencing steps. The sample is immobilised onto specifically designed DNA capture beads, each bead carrying a unique sequence. Each sequence then undergoes an Emulsion PCR (emPCR) step and is amplified on its individual bead. Each bead is then loaded onto a PicoTiterPlate which allows for only one bead per well. The sequencer then flows individual nucleotides in a fixed order across the hundreds of thousands of wells. Addition of one (or more) nucleotide(s) complementary to the template strand results in a chemiluminescent signal

recorded by sequencer. The combination of signal intensity and positional information generated across the PicoTiterPlate allows the sequencing software to determine the sequence of each of the input sample.

One of the drawbacks of 454 sequencing is that the sequencing software uses signal intensity to determine the number of consecutive identical bases in a sequence. When multiple consecutive identical bases are encountered (especially four or more repeated bases) the software cannot reliably interpret the signal intensity (and therefore the number of bases read) which can lead to sequencing errors especially with low complexity sequences.

Recently Solexa/Illumina machines [Bennett, 2004] have further increased the number of reads obtainable from sRNA deep-sequencing runs. This approach allows in excess of 1 million sRNA reads per experiment with a read length of 35nt making it ideally suited for sRNA discovery. This technology again requires 3' and 5' adaptors to be ligated to the sample to be sequenced. The sample is then attached to a special optically transparent plate. Attached DNA fragments are extended and amplified to create an ultra-high density sequencing flow cell with over 50 million clusters, each containing around 1000 copies of the same template. These templates are sequenced using a four colour DNA "sequencing-by-synthesis" technology. This approach means that each base is sequenced individually base-by-base and eliminates the problem that 454 technology has with repetitive sequence.

Although high-throughput techniques have revolutionised sRNA sequencing they

have led to new problems with data analysis. Previously, biologists would often manually work through small lists of sRNAs but with millions of reads now being produced by a single sequencing run the need for computational techniques to process and classify sRNAs has become apparent.

2.3.3 miRNA target validation

As previously described plant miRNAs tend to be highly complementary to their targets and cause mRNA cleavage at a specific position within the target site of the miRNA [Rhoades et al., 2002]. This means that predicted targets can be validated experimentally by a process called 5' **R**apid **A**mplification of **c**DNA **E**nds (5' RACE) analysis. In essence the process allows the sequencing of cleavage products from the mRNA predicted to be targeted by a given sRNA. The sequences can then be aligned to the full length mRNA and, if the mRNA is regulated, then the cleavage products should begin at the precise nucleotide position predicted to be targeted by the sRNA. This method is not applicable to animal target validation as the mRNAs are not usually cleaved. Instead a luciferase assay is required [Clancy et al., 2007, Wang et al., 2007]. As this technique is not relevant to later work presented in the thesis it will not be detailed here.

New methods proposed by Addo-Quaye *et al*. [Addo-Quaye et al., 2008] and German *et al*. [German et al., 2008] allow the entire transcriptome of an organism to be searched for miRNA cleavage products, and, unlike in the traditional 5'RACE method, the miRNA and proposed target do not need to be known before the experiment is

```
miR171                      3’ CUAUAACCGCGCCGAGUUAGU 5’
                               ||||||||||||||||||||
AT4G00150 GACACGTGTCTAGCTCAGGGGATATTGGCGCGGCTCAATCAACAGCTCTCTTCTCCC
cleavage1 ------------------------------GGCTCAATCAACAGCTCTCTTCTCCC
cleavage2 ------------------------------GGCTCAATCAACAGCTCTCTTCTCCC
cleavage3 ------------------------------GGCTCAATCAACAGCTCTCTTCTCCC
cleavage4 ------------------------------GGCTCAATCAACAGCTCTCTTCTCCC
cleavage5 ------------------------------GGCTCAATCAACAGCTCTCTTCTCCC
```

Figure 2.7: Alignment of *Arabidopsis* AT4G00150, Scarecrow-like 6 (SCL6) mRNA to its 5’ RACE cleavage products. miR171 target site is shown along with base pairing between miRNA and target mRNA. 5’ start positions of all cleavage products all begin the cleavage site showing a precise cleavage between positions 10 and 11 of the miRNA. Only the region of the alignment around the cleavage site is shown in this figure.

performed. The groups combine a modified 5’RACE with high-throughput deep sequencing to create libraries that contain 3’ cleavage products of mRNAs. mRNA fragments from the high-throughput sequencing are computationally mapped to a library of mRNAs and potential target sites can be distinguished from random mRNA degradation products using the relative abundance of the high-throughput sequence reads. In their test data, highly abundant reads mapped to the cleavage positions of the mRNAs known to be targeted by miRNAs making genuine targets easily distinguishable from background noise and providing a novel high-throughput approach for the identification and validation of miRNA targets.

2.4 Computational prediction and analysis of sRNAs

Ever since their discovery, groups from all over the world have been hunting for new miRNAs (and homologues of known miRNAs) in a variety of different organisms.

As previously mentioned, miRNAs are derived from a pre-miRNA (see figure 2.1), the structure of which is evolutionarily conserved. Several RNA folding algorithms e.g. [Zuker and Stiegler, 1981, McCaskill, 1990] have been implemented that allow the prediction of such structures. In general, these work using dynamic programming algorithms (see [Eddy, 2004] for review) which allow the prediction of a minimum free energy (MFE) structure for a given input RNA sequence. Examples include `RNAfold` [Hofacker, 2003] and `Mfold` [Zuker, 2003]. In general, single sequence structure prediction can be quite unreliable [Gardner and Giegerich, 2004] but due to the high degree of base pairing in miRNA precursors, predictions of pre-miRNA structures are generally believed to be much more accurate and show greater stability than other non-coding RNAs [Chan and Ding, 2008, Loong and Mishra, 2007b, Bonnet et al., 2004b].

Purely computational prediction of other sRNA classes has largely been ignored due to the fact that they are derived from transcripts that do not have a conserved secondary structure or sequence making it almost impossible to design algorithms for their classification.

2.4.1 miRNA homologue detection

Several approaches have been employed to detect homologues of known, experimentally validated miRNAs (see [Weber, 2005, Legendre et al., 2005, Dezulian et al., 2006, Hertel et al., 2006, Artzi et al., 2008] for details). All use a sequence similarity search such as BLAST [Altschul et al., 1990], which looks either for the mature miRNA or pre-miRNA in a target genome. The secondary structure

of candidate pre-miRNA is then analysed to ensure that the characteristic hairpin structure is conserved. Other more advanced single sequence search methods such as RSEARCH [Klein and Eddy, 2003] which takes into account RNA secondary structure may also be effective for this purpose. Homology searching is relatively straightforward and does not pose any significant computational challenges as those mature miRNA sequences that are conserved are generally highly similar on a sequence level even though the organisms they come from may be evolutionarily very distant.

2.4.2 *de novo* miRNA prediction

De novo miRNA prediction, where novel miRNA candidates are computationally detected from a given input genome sequence without any prior knowledge of their existence, is a much more challenging problem than homologue detection. The difficulty is due to the small size of the sequences to be detected and large size of the search space (the target genome sequence).

Several computational methods have been described for the *de novo* prediction of miRNA sequences [Lai et al., 2003, Lim et al., 2003, Jones-Rhoades and Bartel, 2004, Bonnet et al., 2004a, Wang et al., 2005], many of which are based on assumptions made by Lai *et al*. [Lai et al., 2003] which state that miRNAs are highly conserved between the genomes of related species. In general, these approaches are based on folding sequence windows of a defined length from an input genome and predicting the most energetically stable secondary

structure with algorithms such as RNAfold and Mfold. Resulting secondary structures are then assigned scores based on empirical properties of known miRNA precursor hairpins.

Most methods produce many thousands of candidate miRNA sequences, indicating that such approaches suffer from a lack of specificity. To reduce the number of false positive predictions, many algorithms (e.g. [Wang et al., 2004b, Jones-Rhoades and Bartel, 2004, Bonnet et al., 2004a, Wang et al., 2005, Adai et al., 2005]) employ a conservation rule, i.e. a candidate miRNA is only accepted if a homologue can be found in the genome of at least one other related species. This method of miRNA prediction has been successfully employed to find many novel miRNAs in both plants and animals with a high degree of accuracy.

Although some miRNAs are conserved between closely related organisms, many have now been shown to be specific to individual taxonomic groups [Barakat et al., 2007b, Bentwich et al., 2005, Fahlgren et al., 2007, Yao et al., 2007]. This discovery has exposed the limitations of comparative methods and has led to the need for alternative approaches to miRNA detection. Recently a Support Vector Machine (SVM) based classifier miPred [Loong and Mishra, 2007a] was released. miPred claims to give accurate predictions without the need for cross-species validation and achieves a 84.55% sensitivity and 97.75% specificity on a test set of human pre-miRNAs and pseudo-hairpins (the negative control). An earlier study suggests

that computational folding of sequence windows obtained from the human genome can yield around 11 million hairpin sequences [Bentwich et al., 2005] and given an estimated false positive rate of 2.25% the method would likely give rise to around 250,000 false positive miRNA predictions, a scale of error that would be unacceptable for biologists who need to validate candidate miRNAs experimentally.

siRNAs are derived from primary transcripts that do not have a well conserved secondary structure or sequence properties. This means that siRNA loci cannot be predicted computationally without knowledge of sRNA sequences (such as those produced from high-throughput sRNA sequencing experiments). As such no *ab initio* siRNA detection methods have been published to date.

2.4.3 miRNA target prediction

In plants miRNAs tend to have near perfect complementarity to their target genes and usually bind within the coding region of the mRNA leading to cleavage and degradation of the sequence. Because of the high degree of complementarity between miRNA and target, plant miRNA target prediction is thought to be a relatively straightforward process which relies on simple sequence searches such as BLAST [Altschul et al., 1990] and FASTA [Pearson and Lipman, 1988] to find potential target regions. The regions are then filtered based on a number of rules (e.g. [Allen et al., 2005, Schwab et al., 2005]) such as the number of mismatches between the miRNA and target, as well as the MFE of the miRNA/target duplex. Many plant

targets have been identified computationally in this way and later validated experimentally. It is therefore thought that plant miRNA target prediction is much more accurate than animal target prediction.

In animals the situation is quite different. miRNAs tend to target mRNAs within their 3' untranslated region (3' UTR) and show a low degree of complementarity to their target site. Unlike in plants, the mRNA is not usually cleaved but instead the binding of the RISC along with the miRNA acts to block translation of the mRNA into its protein product. The low complementarity between miRNA and target in animal systems makes it extremely difficult to predict accurately the targets of a given miRNA. Even so, most animal miRNAs do show a greater degree of complementarity to their targets at their 5' ends (positions 2-7 or 8) known as the "seed" sequence [Lewis et al., 2005].

There have been several attempts to address the animal target prediction problem computationally (see e.g. [Enright et al., 2003, Krüger and Rehmsmeier, 2006, Krek et al., 2005]). These methods rely on finding target sequences from a single miRNA input, and employ nucleotide complementarity and MFE calculations to identify miRNA/target duplexes and rank target sites based on these parameters.

Although these methods have been successfully used to predict some real targets, they tend to lack specificity, and can produce many false positives. Different target prediction methods also tend to predict candidate targets that show little or no overlap [Rajewsky, 2006].

As with detecting the miRNAs themselves, comparative genomic approaches have

been successfully used to identify inter-species target site conservation so as to filter out non-conserved target sites (e.g [John et al., 2004, Watanabe et al., 2006]). This form of post processing can improve specificity but can also reduce sensitivity meaning that important targets can be missed.

Other strategies for improving specificity of target predictions include looking at target site accessibility. Many predicted targets are located within regions of 3' UTRs which are highly structured. If the target is part of a very stable secondary structure then it is unlikely to be accessible to the miRNA and therefore should not be considered as a valid target site. Kertesz *et al*. [Kertesz et al., 2007] proposed that target accessibility is a critical factor in microRNA function and target prediction.

2.4.4 ta-siRNA prediction

Although it is not possible to predict TAS loci directly from a genome sequence, the characteristic phased nature of ta-siRNAs make it possible to classify loci based on data derived from high-throughput sequencing as long as the genome sequence of the organism is known.

High-throughput sRNA data can be mapped to a reference genome in order to identify sRNAs phased in 21nt increments that could represent potential TAS loci. Chen *et al*. recently published a method to detect TAS loci from high-throughput sequence data [Chen et al., 2007]. This method relies on a simple statistical test to look at the number of phased and non-phased positions within a cluster. Firstly a 231nt window downstream of the 5' start site and an 231nt antisense window with a 2nt 3'

shift (to mimic the DCL4 cleavage) is produced relative to each sRNA (Figure 2.8A). This region contains contain 21 possible “phased” positions in 21nt increments and 440 possible “non-phased” positions relative to the start site of each sRNA. The number of distinct small RNAs mapping to this region (n) and the number of distinct small RNAs mapped to phased positions (k) are counted. A p-value of obtaining k or more phased small RNAs is calculated from the hypergeometric distribution (Figure 2.8B). Low p-values (p < 0.01) represent good candidates for ta-siRNA producing loci.

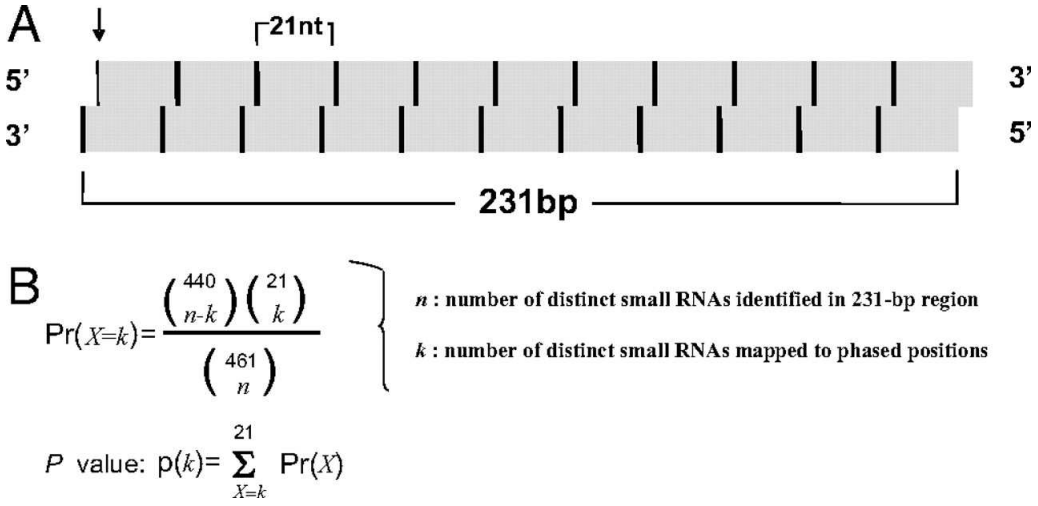

Figure 2.8: Theoretical basis and derivation of the TAS prediction algorithm. A) The vertical arrow indicates the start site for the small RNA used to determine the phased and non-phased positions. 21 phased sites relative to the start site are indicated as black vertical bars. Four hundred forty non-phased sites relative to the start site are indicated as gray. B) Equation based on hypergeometric distribution for statistically evaluating the probability of obtaining k or more phased sRNAs from the genomic fragment defined in A) This figure is taken from [Chen et al., 2007].

Although this method has been successfully used to identify both known and novel TAS loci [Chen et al., 2007] it is not without its drawbacks. First, it is known

that ta-siRNAs are 21nt in length [Allen et al., 2005] and yet the Chen *et al*. algorithm only looks at the sRNA start position on the genome and does not take size information into account. The second main drawback with this method is that it ignores abundance (cloning frequency) information so highly abundant sRNA reads are treated equally to those with a single read count. This makes the algorithm highly susceptible to noisy data, and it is known from analysing published 454 data (e.g. [Fahlgren et al., 2007, Rajagopalan et al., 2006]) that phasing at known TAS loci is often imperfect. However “in-phase” sRNAs are often present at much higher abundance than other sRNA mapping to the loci. It is not known whether this is a result of sRNAs mapping to the locus by chance (and in reality being produced from a different genomic location), a product of incorrect sequencing, or the result of imprecise processing of ta-siRNAs in the cell (at a low frequency).

2.5 Discussion

This chapter has given a brief background on sRNAs and the biological and computational methods currently used in their identification, classification and validation. Over the previous five years there have been huge technological advances, and the advent of the high-throughput sequencing has opened up new possibilities in the sRNA field as well as many other areas of biological research. The advent of these new technologies has however caused problems for biologists who need to process and analyse the large amounts of data produced. With new even higher throughput technologies

on the horizon the need for specialised, easy to use tools for the detection of miRNAs, and other sRNA loci of interest are of utmost importance. These challenges lie in the field of bioinformatics and throughout the rest of this thesis we will present new methods and tools designed to improve high-throughput sRNA analyses, as well as discussing some applications of these techniques to biological datasets.

Chapter 3

miRNA detection in high throughput small RNA sequencing data

This chapter is an adapted and extended version of

Moxon, S., Schwach, F., Studholme, D., Dalmay, T., MacLean, D., Moulton, V. (2008): A toolkit for analysing large-scale plant small RNA datasets *Bioinformatics*. In press.

3.1 Summary

This chapter describes a toolkit created to process high-throughput datasets and produce high quality miRNA and ta-siRNA candidates for follow up experimental validation. Firstly we describe `miRCat`, a tool to identify miRNA sequences in high-throughput sequence sets and benchmark its performance on published data. We then go on to describe a suite of publicly available web-based tools consisting of an online implementation of `miRCat`, a plant sRNA target prediction tool, a ta-siRNA prediction method, and `SiLoCo`, a tool to compare the expression levels of sRNA loci in two different sRNA samples.

3.2 miRCat

As described in Chapter 2, there is a need for a generalised tool to screen high-throughput sRNA data sets for miRNAs. At the time of writing, one such tool `miRDeep` [Friedländer et al., 2008] is available for analysing animal datasets, but no such tools have yet been published for use with plant sRNA data. We now describe such a tool that we call `miRCat` (**miR**NA **Cat**egoriser).

3.2.1 Features

For flexibility, adaptability and future extendability, `miRCat` is comprised of a collection of standalone scripts written in Perl. It takes as input two Fasta format files;

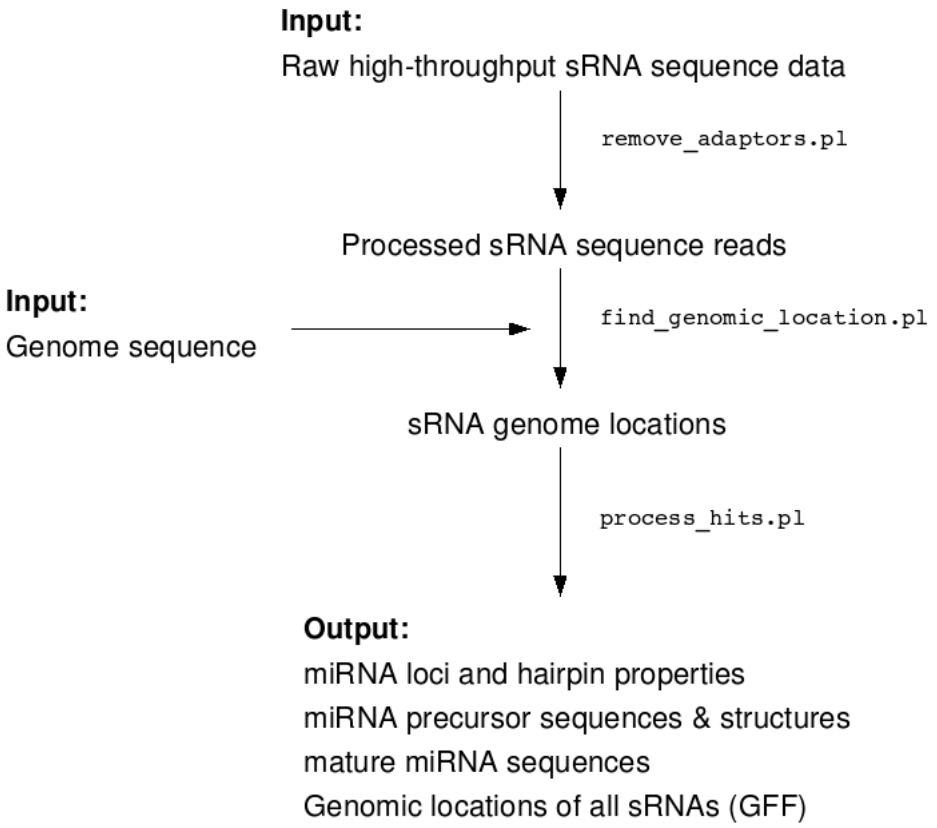

Figure 3.1: Workflow diagram showing inputs and outputs of the `miRCat` pipeline.

one containing the sRNA sequences, and the other the corresponding genome sequence (or set of sequences) from the organism of interest. The sequences are preprocessed to remove adaptor sequences (used in the experimental sequencing of the sRNAs), mapped to the genome, and miRNA candidates are predicted. The output of `miRCat` consists of a comma separated value (.csv) format file (easily imported into any spreadsheet software) containing information about each candidate miRNA. A Fasta file of candidate mature miRNA sequences, and a plain text file containing structure predictions for the precursor miRNAs. Optionally a GFF format file [GFF, 2000]

containing all sRNAs and their genomic locations can be generated. For an overview of `miRCat` see Figure. 3.1.

Adaptor removal

Adaptor sequences are commonly used in high-throughput technologies such as 454 and Illumina and need to be removed before any further analysis is performed. `remove_adaptors.pl` reads in the raw reads generated from the sequencing and prompts the user for a set of adaptor sequences. Sequences that exactly match both 3' and 5' adaptors (in the case of 454 sequence reads) and 3' adaptors (in the case of Solexa/Illumina reads) specified by the user are then removed (see Figure. 3.2). All sequences matching the reverse complement of the adaptors are processed, and then reverse complemented in order to achieve the correct sequence orientation. A full breakdown of the number of sequences in the input set, the number extracted and the number of sequences that did not match both adaptors are then displayed to the user.

Obtaining genomic coordinates

Once adaptors have been removed, genomic coordinates of each sRNA are computed for later analysis. We found existing tools such as BLAST [Altschul et al., 1990] and FASTA [Pearson and Lipman, 1988] to be unsuitable for this task as, although they are fast, they do not guarantee to find all matches to a sRNA on the genome sequence. We instead devised an exact matching technique that reads all sRNAs and their reverse

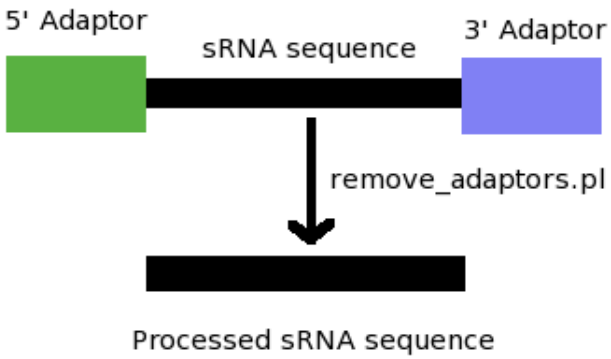

Figure 3.2: Raw sequence reads from high-throughput projects contain 5' and 3' adaptor sequences (green and blue respectively) that must be removed before they can be mapped to the genome. `remove_adaptors.pl` will quickly remove exact matches to adaptor sequences supplied using a simple Perl pattern match.

complements into memory, and then splits the reference genome into fragments that correspond to the lengths of each sRNA in the input set. Each genomic fragment is compared to the sRNA set (and the set of sRNA reverse complements) in memory and if the sequences match exactly then the genomic coordinates of the fragment are recorded. This algorithm was implemented in `find_genomic_location.pl` which reads in a the processed Fasta format file of sRNA sequences and produces a list of coordinates in the user defined genome sequence.

sRNA analysis

`process_hits.pl` is designed to detect candidate miRNA sequences based on their genomic location and putative precursor properties. Firstly, the script computes

```
1/78933-78952(-1)190(2)=TCGGACCAGGCTTCATCCCC
1/537974-537993(1)9948(1)=TTCATCCCCAAATTGATAAC
1/1276554-1276573(1)18080(3)=TCTTGTGTGATGATGTGTCA
1/3785426-3785445(1)46778(1)=AATCATCTTCCACCTCCTTA
1/3825885-3825904(-1)14251(1)=GTTCTTATTAGATGAGGGTA
1/3920120-3920139(-1)36046(1)=TTTTGATGAGCGTTTGAATA
1/3961368-3961387(-1)210(13)=TTGAGCCGTGCCAATATCAC
```

Figure 3.3: Example of output from `find_genomic_location.pl` in the following format: chromosome / start - end (strand) sRNA_accession (sRNA_abundance) = sRNA_sequence.

clusters of sRNA hits (or loci) based on their genomic location. The genome coordinates created in the previous step are read in and sRNAs are clustered based on their proximity to one-another. Neighbouring sRNAs are defined to fall into the same locus if they are within 200nt of one another. 200nt was chosen as the optimum value after looking at the proximity of known *Arabidopsis* miRNA/miRNA* pairs on the genome.

Once genomic loci have been defined, `process_hits.pl` compares properties of sRNA loci to those of known miRNAs. The default settings require 90% of all sRNA hits within a locus to be in the same orientation. Moreover, no more than four distinct, non-overlapping sRNA hits must be present in the locus, and the maximum total length of overlapping sRNAs must not exceed 70nt. If these criteria are satisfied then the locus is further analysed by extracting flanking genomic sequences of varying lengths, both upstream and downstream of the most abundant sRNA in the locus (which is considered to be the miRNA sequence). The resulting sequences are folded with `RNAfold` [Hofacker, 2003] and their secondary structures are checked for a number of properties indicative of miRNA precursors [Ambros et al., 2003].

First the base pairing of the most abundant sRNA (the predicted mature miRNA) within the putative precursor is analysed. The mature miRNA must not exhibit base pairing with itself as no known miRNAs have this feature. If this is the case, then the sequence window is discarded. Otherwise the structure is analysed and, if a hairpin-like configuration (see Figure. 2.1) is found (by analysing the base-pairing and identifying a valid stem and loop region), and is longer than the minimum hairpin length (70nt) then the hairpin is subjected to additional tests. Using default settings (based on recent stringent criteria for miRNA annotation [Jones-Rhoades et al., 2006]) no more that three consecutive unpaired bases are allowed in the 25nt region centred around the predicted mature miRNA and no fewer than 18nt in this 25nt region are allowed to be unpaired.

All structures passing these criteria are filtered based on their adjusted minimum free energy (AMFE) as defined by the user – default -25.0 kcal/mol. The AMFE is calculated by dividing the MFE derived from `RNAfold`, by the length of the potential miRNA precursor sequence and multiplying this value by 100. This calculates the MFE per 100nt of sequence and allows the normalisation of MFEs from sequences of differing lengths [Zhang et al., 2006b].

Sequences fulfilling these rules are then analysed using `randfold` [Bonnet et al., 2004b]. `randfold` shuffles the hairpin sequence, preserving its dinucleotide frequency, and folds the sequence (using `RNAfold`) to obtain its MFE. This process is repeated (`miRCat` uses 100 `randfold` randomisations) and the

MFE of the hairpin is compared with the distribution of MFE values obtained from the shuffled sequences. A p-value is assigned to the hairpin sequence based on its stability compared to the random sample. By default, all hairpins with a p-value of greater than 0.1 are filtered out as they are unlikely to be real miRNA hairpins. This process has been shown to be effective in screening miRNAs [Bonnet et al., 2004b, Loong and Mishra, 2007b].

As multiple sequence windows containing the mature miRNA are checked, the hairpin with the most stable structure (based on its MFE) is chosen as the representative miRNA-precursor sequence.

3.2.2 Testing

`miRCat` has been tested using a variety of published plant and animal datasets taken from the NCBI Gene Expression Omnibus (GEO) repository [Barrett et al., 2007] and shows a good level of sensitivity and specificity. When tested on the Rajagopalan *et al*. wild-type *Arabidopsis thaliana* leaf sample [Rajagopalan et al., 2006] (GEO accession: GSM118373) containing 186,899 sRNA sequences (see Table. A.1), `miRCat` predicts 89 miRNA loci using default parameters. 83 of these predictions are known miRNA sequences (miRNAs from 91 loci were present in this sample), and 6 novel miRNA loci were predicted. This shows a 91.2% sensitivity and, even assuming all the novel predictions are false-positives, this gives a specificity of 99.93% (8362 loci tested). Using the Col-0 leaf sample from Kasschau *et al*. [Kasschau et al., 2007] (GEO accession: GSM154370) which contained 15,833 sRNA sequences (see Table.

A.1), miRCat predicted 51 miRNA loci (miRNAs from 55 loci were present in this sample) i.e. a 92.7% sensitivity. It also detected two potential novel miRNAs in this set.

miRCat has also been adapted to work with animal data. In general, animal miRNA precursors tend to be shorter and less energetically stable than their plant counterparts [Reinhart et al., 2002, Millar and Waterhouse, 2005, Jones-Rhoades and Bartel, 2004]. Animal miRNAs can also be densely clustered as their pri-miRNAs can contain many pre-miRNAs [He et al., 2005, Tanzer and Stadler, 2004, Seitz et al., 2004] so some changes to the method were necessary to allow for this. Based on tests with known animal miRNAs the threshold AMFE was increased to -18.0 kcal/mol, the minimum hairpin length parameter was reduced to 60nt and the maximum distance between neighbouring sRNA hits was reduced to 100nt.

miRCat was tested using sRNAs obtained from a Solexa sequencing run from mouse embryonic stem cells [Calabrese et al., 2007] and predicted 213 miRNAs in this set. Two of the predictions were not annotated and are therefore good candidates for novel mouse miRNAs. Even if these two are false positives it would still mean that the method is above 99.9% specific. All the other predicted miRNAs mapped to either known miRNA loci or new loci of existing miRNAs in miRBase [Griffiths-Jones et al., 2008]. Of the 205 known miRNA loci present in this sample,

164 were correctly identified by `miRCat`, giving an 80% sensitivity using default parameters. Additional filters were included in this analysis and known non-coding RNAs such as snoRNAs taken from Rfam [Griffiths-Jones et al., 2005] were removed from the input dataset. In addition it is likely that further optimisations could be made in order to increase the sensitivity. Full results can be found in Table. A.4.

3.2.3 Applications

We have run `miRCat` on several sRNA datasets including those in [Pilcher et al., 2007, Kasschau et al., 2007, Qi et al., 2006, Rajagopalan et al., 2006, Lu et al., 2006], and it is currently being used to analyse *Arabidopsis thaliana*, *Medicago truncatula*, *Chlamydomonas reinhardtii* and *Gallus gallus* (see A.5) Solexa datasets. We have also combined several publicly available *Arabidopsis* wild-type sRNA datasets obtained using 454 sequencing technology (GEO accessions: GSM118372, GSM118373, GSM149079, GSM154336, GSM154370, GSM257235, GSM118375, GSM121455 and GSM149080) to see whether re-analysing these published datasets using `miRCat` could yield any further novel miRNAs previously missed by other groups.

A total of 1,160,593 sRNA reads were obtained from these datasets after exact matches to known tRNAs and rRNAs from Rfam [Griffiths-Jones et al., 2005], the Arabidopsis tRNA database:

`http://lowelab.ucsc.edu/GtRNAdb/Athal/`

and rRNA sequences obtained from the EMBL nucleotide sequence database

[Kulikova et al., 2007] were removed.

We then used `miRCat` to map the sRNAs to the *Arabidopsis* genome giving a total of 3,416,663 sRNA genomic positions (each sRNA can map to multiple genomic locations).

`miRCat` predicted 131 miRNA loci, 117 of which were previously described miRNAs, and the remaining 14 appear to be strong candidates for novel miRNAs missed in the original analyses by other groups. See Table. A.3 for full results.

3.2.4 Availability

`miRCat` can be downloaded from `http://www.uea.ac.uk/~simonm/miRCat/`. We also implemented `miRCat` as a webserver-based tool to allow users to quickly process their sRNA sets without having to download and install their own local copy of the software, this version is available from `http://srna-tools.cmp.uea.ac.uk/mircat/`.

The command line based implementation of `miRCat` requires installation of Perl version 5.8 `http://cpan.org`, Bioperl 1.4, `http://bioperl.org` or higher, the Vienna package `http://www.tbi.univie.ac.at/~ivo/RNA/` and `randfold` `http://bioinformatics.psb.ugent.be/software.php`. Alternatively users can submit their data for processing via the web-based tool.

3.2.5 miRCat webtool

The web-based implementation of `miRCat` (`http://srna-tools.cmp.uea.ac.uk/mircat/`) (See Figure. 3.4) allows the user to upload a Fasta format file (or zip/gzip compressed Fasta format file) containing their sRNA sequences (up to a limit of 150Mb). The user uploaded sRNAs are first searched against ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) with any exact matches being removed. The remaining sequences are then mapped to a user-selected genome sequence and undergo a series of tests (detailed in Section 3.2.1) in order to classify miRNAs present in the sample. Once the analysis is complete an email is sent to the user with a download link to their results. The output is a compressed archive (zip file) containing a comma separated value (csv) file showing all predicted miRNA sequences present in the input dataset as well as a file containing the secondary structures of predicted precursor miRNA sequences and a Fasta file of predicted mature miRNAs. The user can also opt to create a GFF format file showing the genomic coordinates of all sRNAs on the selected genome sequence.

3.3 UEA sRNA tools server

In addition to `miRCat`, we have implemented several other sRNA analysis tools on the UEA sRNA tools webserver, each of which is described in further detail below.

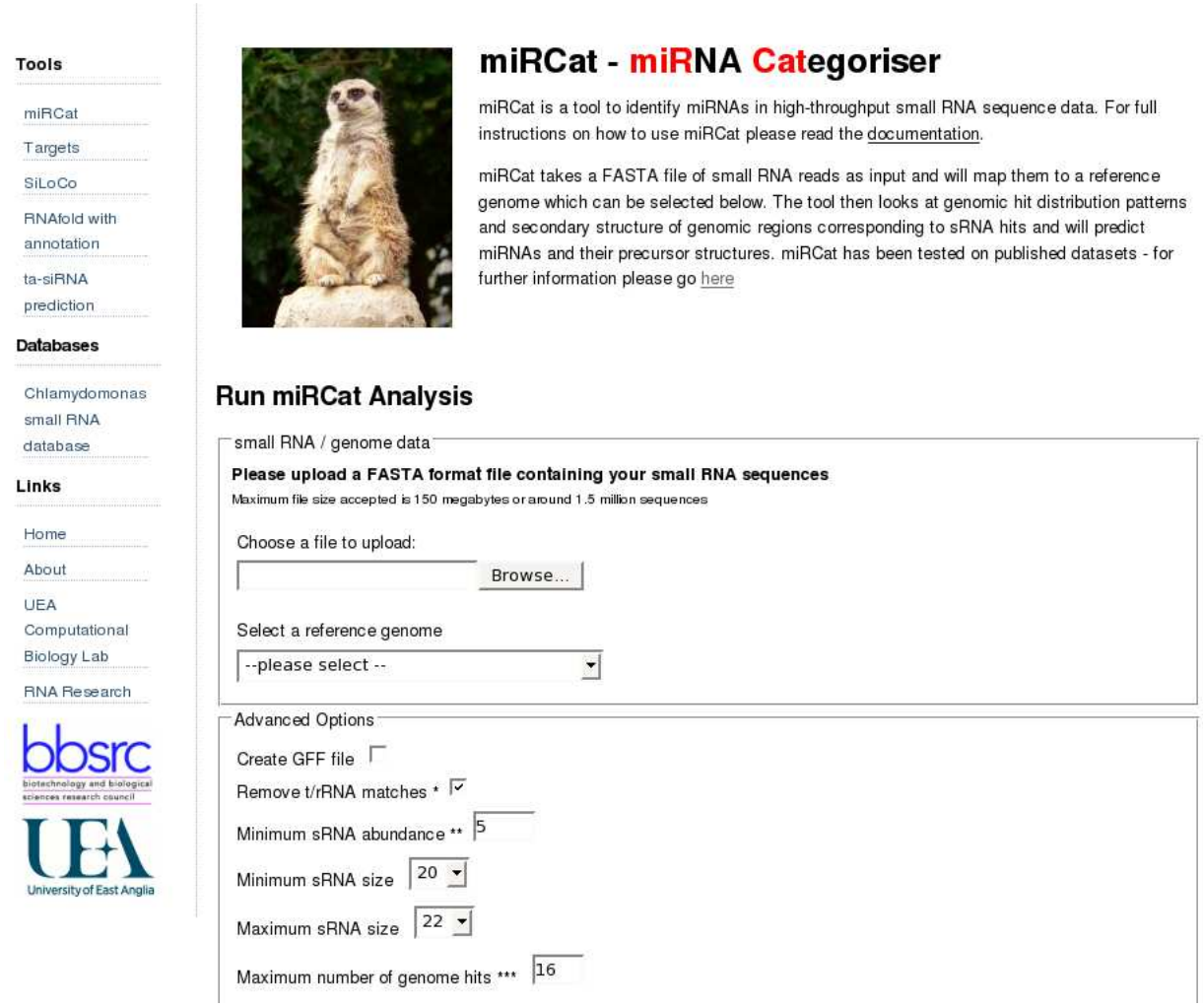

Figure 3.4: Screenshot of the web-interface for the `miRCat` pipeline.

3.3.1 Target prediction

Due to the high degree of complementarity between plant miRNAs and their targets it is possible to accurately predict miRNA-target interactions [Llave et al., 2002, Jones-Rhoades and Bartel, 2004]. We provide a target prediction tool (See Figure. 3.5) which allows users to upload up to 50 sRNA sequences to search for targets in 20 different plant gene sets. The algorithm is based on rules devised by Allen *et al*.

[Allen et al., 2005] and Schwab *et al*. [Schwab et al., 2005] detailed below. Results of target searches along with the full target transcript sequence are emailed to users as an attachment.

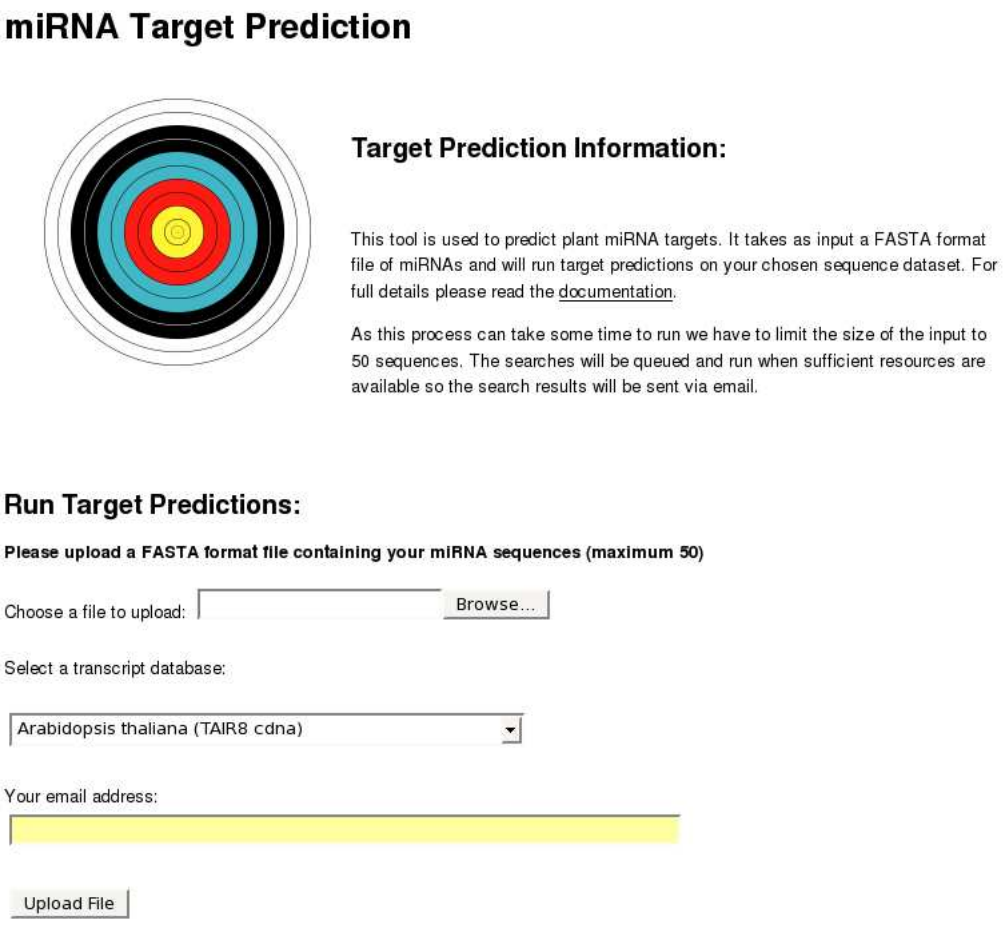

Figure 3.5: Screenshot of the web-interface for the plant target prediction tool.

The target prediction script firstly reads in a set of sRNAs and a flat file sequence database of ESTs or mRNA sequences from an organism of interest (the potential targets). Each sRNA is searched against the potential target set using

FASTA [Pearson and Lipman, 1988]. Any potential matches to the reverse complement of the input sequence then undergo a series of tests based on criteria from [Allen et al., 2005] and [Schwab et al., 2005] to further filter the results. The tests are as follows:

- **Number of mismatches between miRNA and target:** Up to four mismatches between the miRNA and potential target are allowed. G-U base pairs are counted as 0.5 mismatches and all others as 1 mismatch.
- **No more than one bulge in the miRNA/target duplex**
- **No more than two adjacent mismatches**
- **No adjacent mismatches in positions 2-12 (5') of the miRNA**
- **No mismatch in positions 10-11 of the miRNA**: This is due to the fact that cleavage usually occurs at this point and base pairing is always conserved here in known plant miRNA/target interactions.
- **No more than 2.5 mismatches in positions 1-12 of the miRNA**: Generally the miRNA/target duplex is more stable towards the 5' end of the miRNA therefore it is less tolerant of mismatches in these positions.
- **MFE of miRNA/target heteroduplex $\geqslant$ 74% of the perfect (homo-)duplex**: The MFE ratio is calculated by first finding the MFE of the miRNA/target duplex with RNA`cofold` [Hofacker, 2003] and dividing this number by the MFE of the

miRNA bound to its perfect complement. The ratio must be 0.74 or higher and fulfil all other criteria in order to be considered as a target.

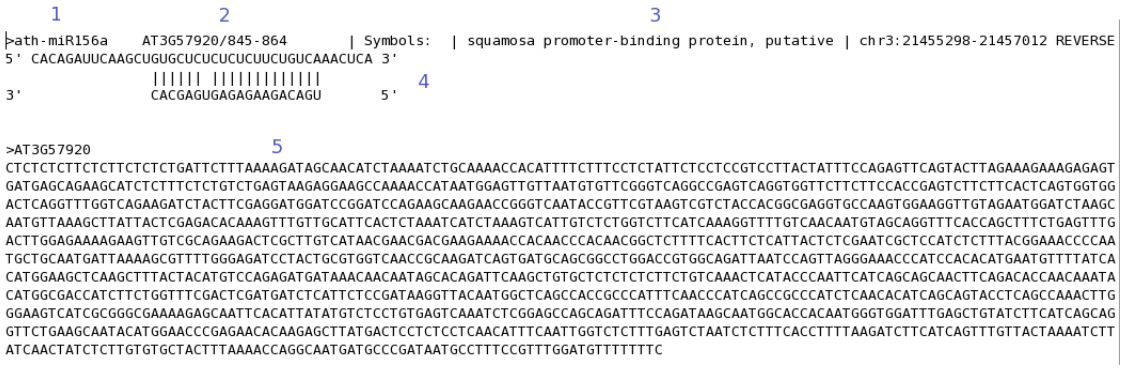

Figure 3.6: Example of the output from the target prediction tool. 1. shows sRNA ID/accession. 2. shows target transcript ID/accession and start-end position of the target site. 3. shows any information/annotation this sequence may have. 4. shows the alignment of the miRNA (bottom sequence) to the target site (top sequence). 5. shows the full sequence of the predicted target.

3.3.2 trans-acting siRNA prediction

To allow for the rapid screening of sRNA datasets for ta-siRNAs, we have made available a modified web-based implementation of the algorithm proposed by Chen *et al*. (See Figure. 3.7) [Chen et al., 2007]. This first maps an input set of sRNAs to a selected genome using `PatMaN`, a recently released program which allows extremely fast exact matching of short sequences to large databases (or genomes) [Prüfer et al., 2008], and then identifies potential ta-siRNA loci as described in section 2.4.4. A p-value is then assigned to each locus and those loci where p is below a user-defined cutoff are classified as candidate ta-siRNA producing locus.

Unlike in the Chen *et al*. implementation we only map 21nt sequences to the genome. All sRNAs of other lengths are removed from the analysis as ta-siRNAs are composed solely of 21nt phased sequences. Chen *et al*. could not do this in their original analysis as it was based on MPSS data so that the size information was not available (all MPSS derived sequences are 17nt long). The other difference in our implementation is that we discard all sequences that were only cloned once in the sequencing experiment. This is due to the fact that single read sequences in high-throughput datasets are unlikely to represent real ta-siRNAs, since, based on observations from *Arabidopsis* datasets, these tend to be found in much higher abundance. Removing single reads also improved the accuracy of the algorithm when run using *Arabidopsis* 454 leaf data from [Rajagopalan et al., 2006] (GEO accession: GSM118373) and [Kasschau et al., 2007] (GEO accession: GSM154370).

The tool was tested using the GSM118373 *Arabidopsis* leaf sRNA dataset taken from Rajagopalan *et al*. [Rajagopalan et al., 2006], where the tool predicted eight ta-siRNA producing loci. Four of these were known TAS loci and two were pentatricopeptide repeat (PPR) gene loci also predicted by other groups as potential TAS loci [Howell et al., 2007, Chen et al., 2007]. The remaining two candidates could potentially be novel ta-siRNA producing loci which need to be further tested using laboratory based methods.

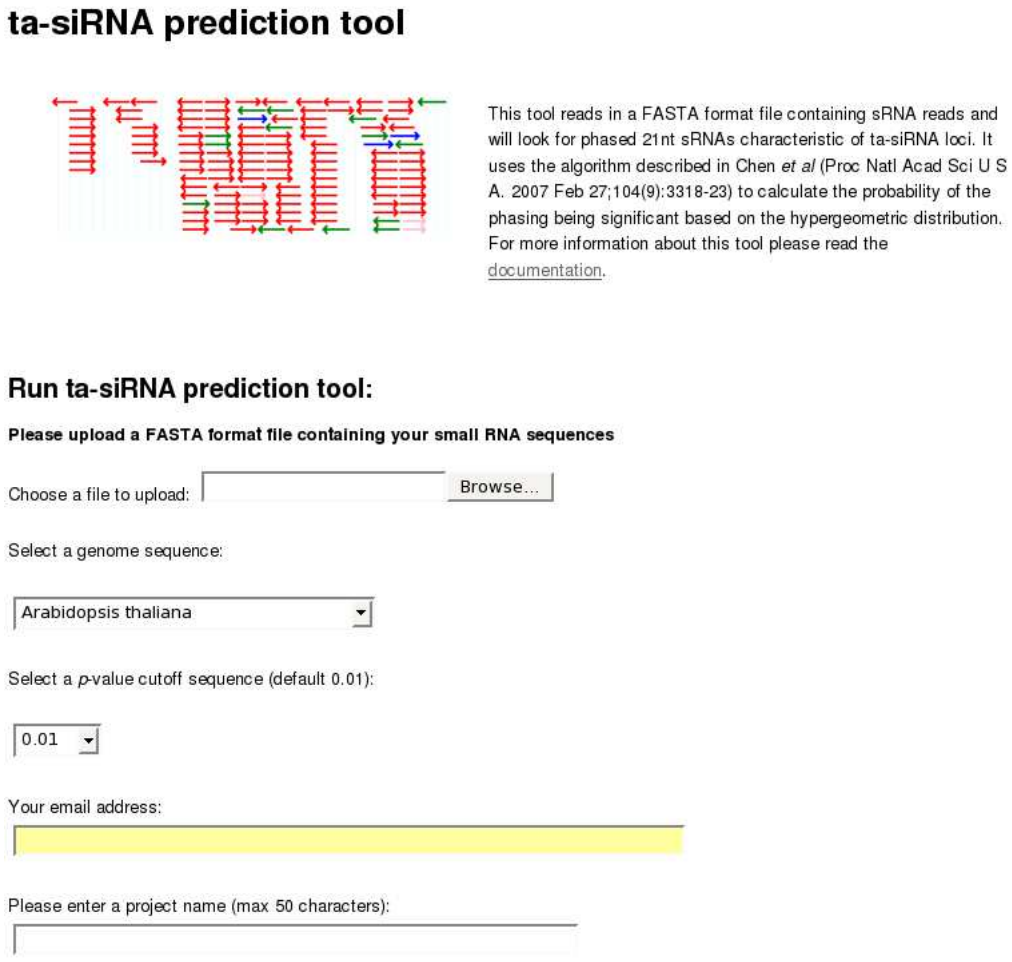

Figure 3.7: Screenshot of the web-interface for the ta-siRNA prediction tool.

3.3.3 Other tools

Two additional tools were implemented by Dr. Frank Schwach as part of the web-tools package. We give a brief overview of both below for completeness.

Table 3.1: Results from ta-siRNA prediction on GSM118373 dataset

Chromosome	Start position	End position	# sequences	# phased sequences	p-value	Locus information
1	18553066	18553317	15	5	2.671539e-04	TAS1b
1	23303204	23303455	7	3	2.545945e-03	PPR repeat gene
1	23423543	23423794	8	3	3.953600e-03	PPR repeat gene
2	11729024	11729275	12	5	7.707355e-05	TAS1a
2	16544727	16544978	10	5	2.603156e-05	TAS1c
2	16546892	16547143	25	7	4.047208e-05	TAS2
3	14214070	14214321	28	5	5.909308e-03	Unannotated locus
3	1970347	1970598	3	2	5.777746e-03	AT3G06435.1

”Chromosome” shows *Arabidopsis* chromosome number; ”Start position” shows start position of predicted TAS locus; ”End position” shows end position of predicted TAS locus; ”# sequences” shows number of unique sequences in predicted TAS locus; ”# phased sequences” shows number of unique sequences that are in phase in the predicted TAS locus; ”p-value” shows the p-value assigned to the TAS locus; ”Locus information” shows any annotation (taken from TAIR [Swarbreck et al., 2008]) for this locus (this column is added manually here for illustration purposes and is not present in the actual output from the tool).

RNA folding and annotation

It is often difficult to visualise where sRNA sequences map to in a longer sequence, so we offer a tool which allows a user to input an RNA sequence such as a miRNA precursor and several sRNA sequences, such as a predicted miRNA and miRNA* sequence. The tool will then fold the longer RNA using RNAfold and mark-up the sRNAs on the larger structure. This can be especially useful for visualising the output from `miRCat` results and also allows users to download publication quality pdf files of their RNA secondary structures. Figure 2.1 was created using this tool, the miRNA and miRNA* sequences are clearly highlighted allowing the user to immediately see the locations of the sRNA in the overall secondary structure.

SiLoCo locus comparison

High-throughput sequencing can be used to compare sRNA expression profiles under varying conditions or between mutants and wild type to gain insights into the biogenesis and function of sRNAs. Plant sRNA populations are highly complex and a simple sequence-by-sequence comparison would not give an accurate picture. To obtain meaningful profiles, sRNA sequences must therefore be grouped into loci of origin and the repetitiveness of sRNA matches to the genome must be taken into account.

`SiLoCo` identifies sRNA loci on plant genomes from two sRNA datasets, which can be uploaded by the user and/or selected from publicly available datasets. `SiLoCo` maps sRNA sequences to the genome and weighs each sRNA hit by its repetitiveness in the genome. Loci are defined as described previously in [Molnár et al., 2007, Mosher et al., 2008] by a minimum number of sRNA hits to a region and a maximum "gap", i.e. absence of sRNA hits, between them. Hit counts are normalised to the total number of genome-matching reads in each sample to make them comparable. For each locus, the log2 ratio and the average of the normalised sRNA hit counts are calculated and ranked independently. A sum of the two ranks is also provided and the results can be downloaded as a csv-formatted file. Sorting the list of loci by the rank sum in a spreadsheet program is an easy way of finding the best candidates for differentially expressed loci where sRNA abundance differs greatly at a high overall expression level. Hyperlinks to some public genome browsers can also be included in the result file (Figure. 3.8 shows a genome browser view of a differentially expressed

locus).

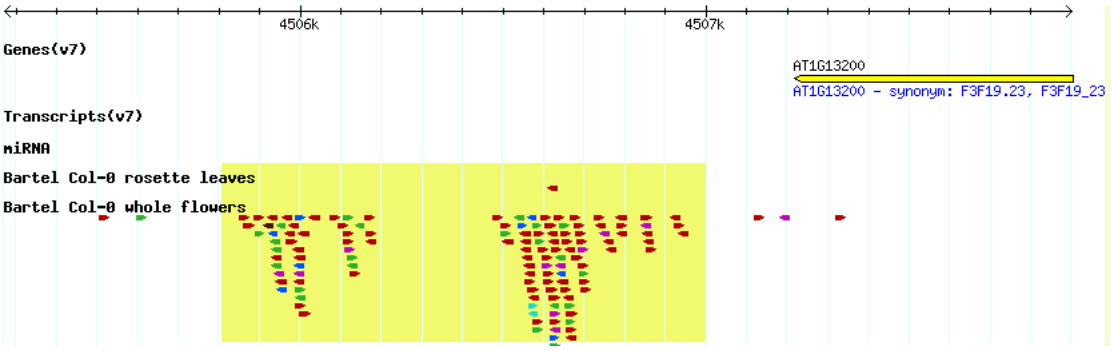

Figure 3.8: SiLoCo candidate locus showing differential expression: The highlighted region (yellow) represents a predicted sRNA producing locus from the SiLoCo tool. Two tracks are visible on the genome browser one from a flower sample (bottom), showing many sRNA hits (coloured arrows) and one from a leaf sample top (showing only a single sRNA hit).

3.4 Discussion

In this chapter we have introduced several new tools to analyse high-throughput sRNA datasets, as well as outlining a web-based toolkit allowing researchers to freely access these resources for use in their own projects. One of these tools, `miRCat`, has successfully been applied to plant and animal high-throughput sRNA data and shows a high level of accuracy on such sets. In the next two chapters, we will describe some in-depth applications of these tools to new high-throughput datasets.

Chapter 4

Identification of novel small RNAs in tomato (*Solanum lycopersicum*).

This chapter is an adapted and extended version of

Pilcher, R., Moxon, S., Pakseresht, N., Moulton, V., Manning, K., Seymour, G., Dalmay, T. (2007): Identification of novel small RNAs in tomato (*Solanum lycopersicum*). *Planta* **226**(3):709-17.

4.1 Summary

The work presented in this chapter was carried out whilst the `miRCat` software described in Chapter 3 was under development. The data analysis helped in the refinement of parameters and rules of the core algorithm in `miRCat` as well as providing feedback from biologists concerning features such as the output format of results.

4.2 Background

Recent work has illustrated the important role that sRNA mediated regulation of transcription factors plays in key developmental processes, such as shoot apical meristem patterning, leaf morphogenesis and flower development [Llave et al., 2002, Palatnik et al., 2003, Aukerman and Sakai, 2003, Baker et al., 2005, Williams et al., 2005]. *Arabidopsis* is the main plant model for studying these processes, although it is different from some species in that upon seed maturation and ripening, it develops a dry dehiscent fruit, similar to that found in cereals and legumes. Other flowering plants have evolved different methods of seed dispersal, such as the animal dispersal of seeds encased within a fleshy fruit. Tomato (*Solanum lycopersicum*) is the model system for studying the biology of climacteric (displays an ethylene burst prior to ripening) fleshy fruit development. The fruit development/ripening process involves many structural, physiological and biochemical changes in the fruit including periods of rapid cell division and expansion, changes in carotenoid accumulation, texture and sugar and

acid content [Giovannoni, 2004] for a review see [Carrari and Fernie, 2006]. MADS box transcription factors have already been associated with tomato fruit ripening with the identification of the ripening-inhibitor gene (LeMADS-RIN) [Vrebalov et al., 2002], an SBP-box transcription factor shown to be necessary for normal ripening in tomato [Manning et al., 2006] and putative transcription factors have also been correlated with ripening and ripening inhibition [Fei et al., 2004].

Tomato fruit development and ripening is a tightly coordinated, highly regulated process, aspects of which are distinct from the events that result in the dry seeds of cereals, legumes and other flowering plants. Given the substantiated role of sRNAs in the regulation of developmental pathways in many other species, particularly in the regulation of transcription factors, we hypothesised that sRNAs may play a key regulatory role in fleshy fruit development and that some of these sRNAs maybe unique to tomato. In this chapter we report the cloning and expression of several known miRNAs from tomato fruit and the identification of 12 novel sRNAs that are not present in *Arabidopsis*, one of them showing a fruit specific expression pattern.

4.3 Materials and methods

4.3.1 Collating and annotating tomato genomic and EST sequences

As there is no complete tomato genome sequence currently available it was necessary to produce a sequence set that included as much information as possible, so that the likelihood of finding matches to the candidate sRNAs could be

maximised. Three sources of data were used to construct the tomato sequence database (TSD): Expressed sequence tag (ESTs) (build: Lycopersicon Combined #3) and BAC (bac.v15.seq) sequences taken from the SOL Genomics Network [Mueller et al., 2005] and annotated genes taken from the EMBL sequence database [Cochrane et al., 2006].

The orientation of many of the tomato ESTs was found to be unreliable and in order to be able to distinguish between potential sources and targets of sRNAs, each sequence in the tomato EST set was blasted against an annotated *A. thaliana* gene set. Any hits with an expected value below 10^{-8} were classified as matches and if the tomato sequence matched the reverse complement of the *Arabidopsis* sequence then it was reverse complemented to correct the orientation. We also found that the annotation of many tomato ESTs did not correspond to the latest *Arabidopsis* annotation and therefore we updated this information.

4.3.2 Cloning of small RNAs

Total RNA was isolated from the pericarp tissue of mature green (39 days after pollination tomato (*S. lycopersicum*) fruit (cv. Ailsa Craig; Horticulture Research International, Warwick, UK). sRNA 19-24nt in length were size fractionated, cloned and sequenced as described in Rathjen *et al*. [Rathjen et al., 2006], from two independent RNA samples (A and B).

4.3.3 Analysis of small RNA sequences

Small RNA sequences greater than 17nt in length were extracted from surrounding adaptor sequences, and sequences containing unassigned nucleotides were removed from each set. To establish a non-redundant set, identical sequences were removed from each cloning. These two sets of sequences were compared to each other and sequences present in both sets, either identical or highly similar to each other (97% similar over 90% length), were identified. The two non-redundant sets, from clonings A and B were also combined, and a new non-redundant set of sequences was derived (C). Sequences from set C were compared to the TSD and exact matches were found using an exact matching algorithm. sRNAs identified either through comparison of sets A and B, or by comparing set C to TSD were filtered by removing degradation products of rRNAs, tRNAs and viral siRNAs. Homologues of known miRNA were also identified at this stage.

4.3.4 Prediction of secondary structures

Sequences that precisely matched the tomato genome, for which sequence flanking the sRNAs could be obtained were analysed by predicting secondary structure. Flanking sequences up to 350nt 5'and 3' of each match were extracted and folded using `RNAfold` from the Vienna RNA package [Hofacker, 2003]. The predicted structure associated with each sRNA sequence was then extended as far as possible whilst maintaining a valid hairpin structure. Any valid hairpins of greater than 75nt with a

MFE of lower than -20 kcal/mol were accepted as being potential miRNAs. These sequences were then analysed using `randfold` [Bonnet et al., 2004b] which assigned a p-value to each of the hairpins based on the probability of the structure occurring by chance.

4.3.5 Northern-blot analysis

Total RNA was extracted from tomato leaf (cv. Ailsa Craig) and whole tomato fruit (cv. Ailsa Craig, mature green stage) as described in Dalmay *et al*. [Dalmay et al., 1993]. Ten micrograms of each total RNA sample was resolved on a 15% denaturing polyacrylamide gel and transferred using a semi-dry electroblotting apparatus to Zeta-probe membrane (Bio-Rad). Four hundred and fifty-nine microgram of total RNA extracted from mature green fruit (as described above) was enriched for sRNAs using the miRVANA kit (Ambion). This sRNA fraction was blotted as described above and these membranes were used to detect sRNAs seen only faintly with 10μg total RNA. Membranes were hybridised overnight at 37°C in ULTRAhyb-Oligo hybridisation buffer (Ambion) with γ-ATP labelled oligonucleotides complementary to each sRNA. Membranes were stripped of probe by incubation in a solution of 10 mM Tris/HCl (pH 8.5), 5 mM EDTA, 0.1% SDS at 95°C for 10 min. Removal of probe was assessed by exposing the membrane for as long as any previous exposure.

4.3.6 Identifying potential *Arabidopsis* homologues of tomato sRNA

Homologues of sRNAs present in both clonings or with an appropriate secondary structure that demonstrated accumulation of 19-24nt RNA species, were searched for in the *Arabidopsis* genome [Rhee et al., 2003] (version TAIR6_cDNA_20051108) permitting a maximum of three mismatches. Where putative homologues were identified, oligonucleotides complementary to these sequences were hybridised to membranes containing 10μg total RNA from both *Arabidopsis* siliques and leaf tissue (Col-0). As a positive control, membranes also contained an oligonucleotide corresponding to the sense orientation of the predicted *Arabidopsis* short sequences. Northern-blot analysis was carried out as described above.

4.4 Results

4.4.1 Tomato genomic and EST sequences

To analyse our results we established a TSD consisting of 34,988 sequences; comprising 30,576 EST sequences, 87 BAC sequences and 4,325 EMBL sequences. The 30,576 EST sequences from the SOL Genomics Network represented a unigene set derived from non-redundant EST contigs and singletons. We noticed that the orientation of some of the unigene sequences was not correct and decided to examine the whole database. It is important to determine whether an sRNA is derived from an EST sequence (same sense as EST) or can potentially target it (complementary to

EST). 5,407 (17.7%) unigenes were found to be in the antisense orientation and were replaced by the reverse complement of the unigene sequence. In addition, based on the *Arabidopsis* annotation we were able to annotate another 5,642 (18.4%) unigenes that had not previously been assigned a function. Although sequence redundancy had been addressed within the unigene EST dataset, the infancy of the tomato genome sequencing project meant that redundancy could be observed both within the BAC sequences and between the unigene, BAC and EMBL datasets.

4.4.2 Identifying expressed sRNAs

The complete genome sequence of tomato is not available, which presents a considerable limitation in the bioinformatic analysis of cloned sRNA. We have cloned sRNA from tomato fruits in two independent experiments because we hypothesised that if a sRNA was detected in two independent samples then it is more likely to have been specifically produced rather than being a random fragment originating from a longer RNA. Two independent direct clonings of sRNA (A and B) from mature green tomato fruit (cv. Ailsa Craig) identified 2,107 and 1,911 sRNAs greater than 17nt in length, respectively. The flowchart illustrating the analysis of the sequences is shown in Figure. 4.1.

Sequences identical in both composition and length from each set were removed, resulting in non-redundant sets of 1,760 and 1,657 sRNA for clonings A and B, respectively. The majority of cloned sRNAs both in sets A and B were 21nt in length

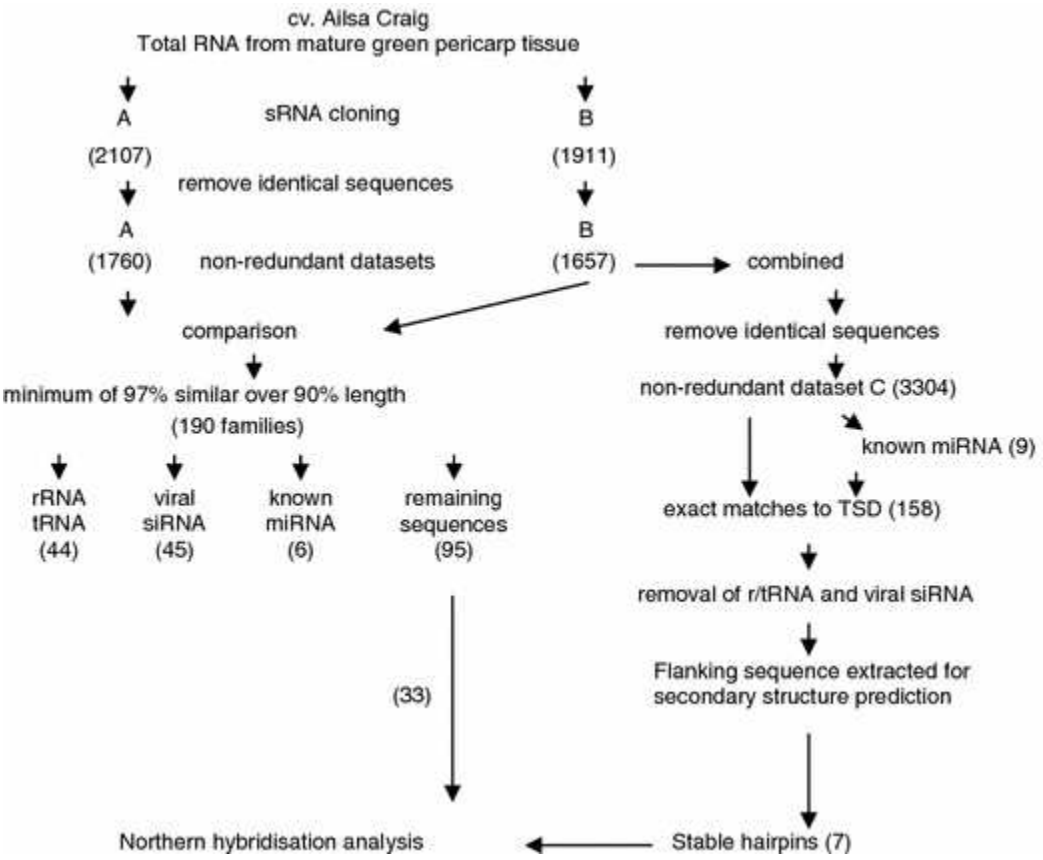

Figure 4.1: Analysis of cloned sRNA sequences. The figure shows the bioinformatic analysis of sequence sets A, B and C leading to the selection of sRNA sequences for Northern-blot analysis. Numbers of sequences pertaining to each stage are shown in *parentheses*.

(27 and 22%, respectively; Figure. 4.2). Comparison of non-redundant set A with non-redundant set B identified 190 families of sRNA that were 97% similar in sequence across 90% of their length.

Within these 190 groups of sRNA, 44 families were identified as breakdown products of ribosomal RNA and transfer RNA, and 45 originated from tobacco mosaic virus. These families therefore were not analysed further. Six known miRNAs occurred in both clonings; miR159, miR162, miR164, miR168, miR171, and miR482

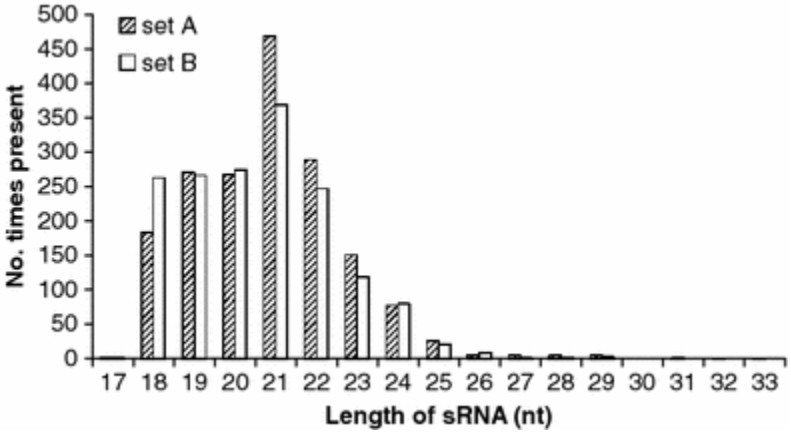

Figure 4.2: Size distribution of cloned sRNA. Frequency of sequences greater than 17nt in length, present in non-redundant sets A and B was plotted against the length of the cloned sequences. The most frequently cloned sRNA size in each dataset was 21nt RNA.

(Reinhart et al. 2002; Lu et al. 2005a, b). Accumulation of tomato miRNAs homologous to these known miRNAs was confirmed by Northern-blot analysis (Figure. 4.3 a). Only miR171 demonstrated differential expression between leaf and fruit tissues, with greater expression in leaf than fruit. From the remaining 95 sRNA sequences 33 were randomly selected for Northern-blot analysis. Six of these sRNAs were expressed and accumulated as 19-24nt species (sRNA3, 4, 5, 7, 8, 9 Figure. 4.3b). Two additional sRNAs that were difficult to detect using 10μg of total RNA, could be clearly detected accumulating as 19-24nt RNA in the purified sRNA fraction from tomato fruit (sRNA 1 and 2, Figure. 4.3b). Probes against nine sRNAs gave signal on the membranes but significantly higher than the 19-24 zone and the remaining 16 probes did not give any detectable signal. These 16 sRNAs that could not be detected by Northern-blot analysis may still accumulate as 19-24nt RNA, but may be expressed at a very low

level and/or expressed in very few cells.

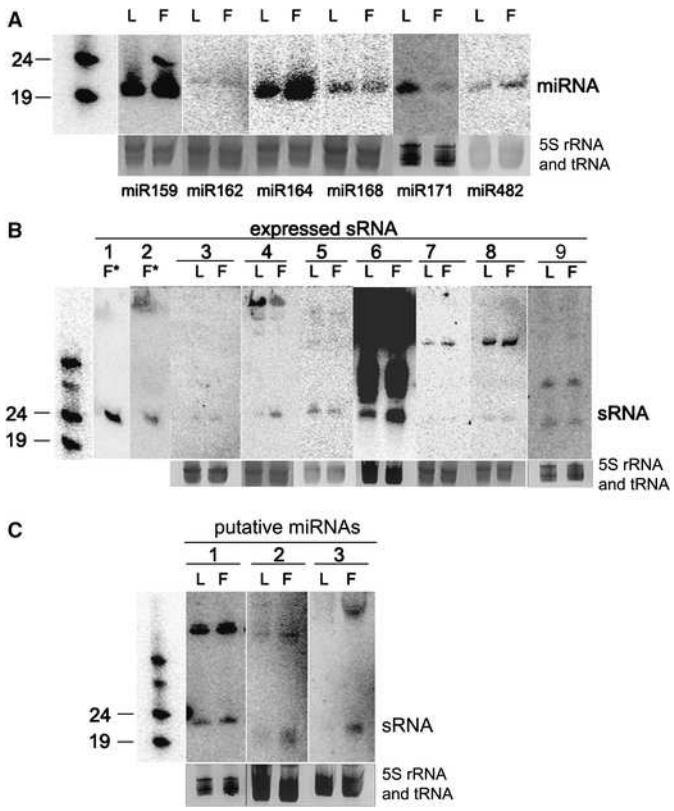

Figure 4.3: Northern-blot of cloned small RNAs. 10μg of leaf (*L*) and mature green fruit (*F*) tissue was loaded onto each gel, unless otherwise stated. Size markers (19 and 24nt) are shown in the first lane of each panel. **A)** Expression of tomato homologues of known miRNAs. Oligonucleotides complementary to tomato miRNAs homologous to known *Arabidopsis* miRNAs were hybridised to membranes to validate their expression. **B)** Expression of sRNAs. Oligonucleotides complementary to cloned tomato sRNAs were hybridised to membranes to validate their expression. The first *two lanes* (marked with *) contained small RNA fractions purified from 459μg of total RNA to detect the expression of sRNA1 and sRNA2 that initially gave very faint signals with membranes containing 10μg total RNA. **C)** Expression of putative tomato miRNAs. Oligonucleotides complementary to cloned sRNAs with predicted stem-loop structure precursors were hybridised to membranes to validate their expression. miRNA3 showed fruit specific expression.

4.4.3 Identifying putative miRNAs

Combining the non-redundant sequences of sets A and B resulted in set C comprising of 3,304 sRNA sequences. Searching for known miRNAs among these sequences identified nine conserved miRNAs (miR159, miR160, miR162, miR164, miR166, miR168, miR171, miR408, miR482 [Reinhart et al., 2002, Sunkar and Zhu, 2004, Lu et al., 2005b, Lu et al., 2005a]. Comparison of set C to the TSD identified seven sRNAs that had no homology with rRNA, tRNA or viral RNA and were predicted to possess an appropriate secondary structure with flanking sequences (hairpin longer than 75nt and MFE < -20 kcal/mol; Figure. 4.4). Two of these sRNAs were identified as tomato homologues of known miRNAs (miR171 and miR168).

Northern-blot analysis of the remaining five sRNAs demonstrated the accumulation of four sRNAs as 19-24nt RNA, with one exhibiting fruit specific expression (Put-miRNA1, 2 and 3 on Figure. 4.3c and sRNA6 on Figure. 4.3b). A range of highly expressed, larger RNA products were observed upon Northern-blot analysis of one of these sRNAs (sRNA6, Figure. 4.3b). This hybridisation pattern is not characteristic of known miRNAs detected to date and therefore this sRNA, despite its predicted stem-loop structure was classed as a sRNA. The remaining three sRNAs (Put-miRNA1, 2 and 3) are novel putative miRNAs since they were cloned as sRNAs, their size and expression was confirmed by Northern-blot analysis, and their predicted precursor can be folded into a hairpin structure [Ambros et al., 2003].

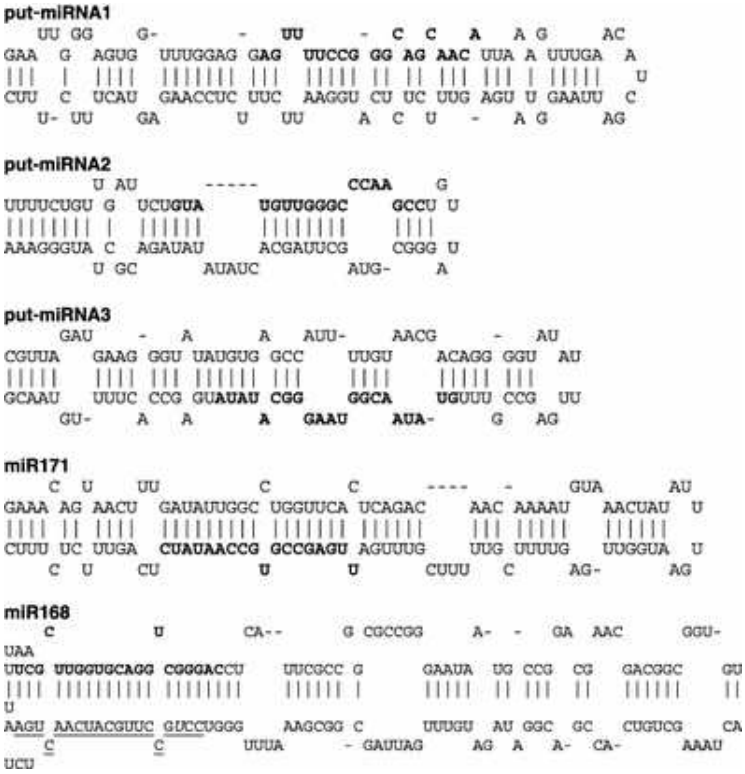

Figure 4.4: Predicted secondary structures of precursors containing sRNAs with exact matches to the TSD. Stem-loop secondary structures greater than 75nt in length and MFE < -20 kcal/mol are shown for three putative tomato miRNAs that were validated by Northern-blot analysis and two cloned tomato homologues of known miRNAs; miR168 and miR171. The sequence of the cloned sRNA is shown in bold text. We also cloned miR168* that has not been identified in *Arabidopsis* (underlined nucleotides). Two other predicted stem-loop structures are not shown because one of the cloned sRNA was not detected and the other (sRNA6) gave extra bands by Northern-blot analysis therefore these are not classed as putative miRNAs.

4.4.4 The new sRNAs are not conserved in *Arabidopsis*

We investigated whether the 12 new sRNA sequences are tomato specific sequences or if they are present in *Arabidopsis*. Searching the *Arabidopsis* genome revealed no exact matches to the 12 new sRNAs. However, it is known that there can be a few mismatches within a miRNA family even within the same plant species and therefore partial matches to the 12 new sRNAs were searched for in the *Arabidopsis*

genome, allowing one to three mismatches. Matches were found for nine new sRNAs (six sRNAs and three putative miRNAs) but no hairpin structures could be predicted in *Arabidopsis* using flanking sequences. Northern-blot analysis showed that while the oligonucleotides complementary to the sRNAs hybridised to control DNA oligonucleotides identical to the predicted *Arabidopsis* homologues, no accumulation of 19-24nt species was seen in the *Arabidopsis* leaf and silique tissues (data not shown), indicating that these sRNAs are not expressed in *Arabidopsis*.

4.5 Identifying novel sRNAs

We identified 12 novel sRNAs in tomato validated through their cloning and expression as 19-24nt RNAs. A key factor in classifying these sRNAs is the prediction of their hairpin precursor, a process that required flanking sequences surrounding the sRNAs. As the sequencing of the tomato genome has only recently begun, we collated the existing available tomato sequence data. The largest proportion of sequence data (30,576 sequences) came from the tomato unigene set, which is estimated to represent approximately 19,800 genes of an estimated 35,000 genes in the 950 Mb tomato genome (Van der Hoeven et al. 2002). Even with the addition of BAC and EMBL sequences, the absence of a large amount of genome sequence for tomato was a limiting factor in the identification of new miRNAs from our cloned sRNAs. A previous study [Zhang et al., 2006a] has shown that the number of miRNA identified within an EST database is linearly related to the number of ESTs, with approximately

10,000 ESTs containing a single miRNA.

In an attempt to circumvent this lack of sequence data, we first compared the sRNA sequences derived from two independent clonings, hypothesising that if a sRNA was present in both clonings then it was more likely to have arisen as the product of a DICER-like gene, rather than a random product of RNA degradation. We expected to clone a high number of rRNA degradation products from the fruit tissue, because microarray data has shown the largest group of fruit upregulated genes are putative ribosomal protein encoding genes [Carbone et al., 2005], but found only 14.7% of our non-redundant set C could be annotated as rRNA or tRNA derived sequences. This percentage was slightly higher among the 190 sRNA families that were present in both sets A and B (23.3%), although we expected that the ratio of randomly generated RNA fragments would be lower in the sequences found in two independent experiments. One possible explanation is that these sRNAs are produced by DCL3 that is involved in the generation of siRNAs from repeat regions including 5S rRNA genes [Pontes et al., 2006].

Arabidopsis candidate sequences partially matching nine sRNAs identified in this work were tested by Northern-blot analysis. None of these potential *Arabidopsis* homologues accumulated as a 19-24nt RNA species in leaf or silique tissues. This indicates that these novel tomato sRNAs are not conserved between the *Arabidopsis* and tomato genomes.

4.6 Classification of non-conserved sRNAs

Eight out of the nine sRNAs detected by comparing sequence sets A and B cannot be further classified, as no exact match was found in the tomato genome and hence no secondary structure could be predicted. Without the ability to predict a secondary structure, we cannot determine if the eight sRNAs have arisen from perfect double stranded RNA (siRNA) or stem-loop secondary structures (miRNA). Since heterochromatin and primary nat-siRNAs are slightly larger (24nt) than mi-, ta-si or secondary nat-siRNAs (21nt) the new sRNA are more likely to belong to one of the previous classes (Figure. 4.3b). Another sRNA (sRNA6) was also classified as sRNA, despite its stem-loop secondary structure because of its atypical hybridisation pattern (Figure. 4.3b).

The available genome information provided flanking sequences for 158 sRNAs and three of these fulfilled the criteria of miRNAs described by Ambros et al. (2003): (1) they could be folded into a stem-loop structure (Figure. 4.4); (2) they did not derive from ribosomal or tRNA; (3) they accumulated as 19-24nt RNA species (Figure. 4.3c) and (4) they were cloned as 19-24nt RNA species. Therefore a few years ago these three short sequences would have been identified as bona fide miRNAs. However, during the last few years a huge complexity of plant sRNAs was revealed, opposite to animal short RNAs [Jones-Rhoades et al., 2006]. Tens of thousands of sRNAs have been cloned from *Arabidopsis* tissues [Lu et al., 2005a, Lu et al., 2006, Rajagopalan et al., 2006], many of them can be detected by Northern-blot analysis

and hundreds of thousands of non-miRNA genomic sequences can be folded into stem-loop structures [Jones-Rhoades et al., 2006]. Conserved miRNAs can be identified with great confidence but it is more difficult to classify non-conserved sRNAs because they could be siRNAs that have precursors with stem-loop structure by chance (due to the very large number of potential stem-loop structures in any genome). Jones-Rhoades *et al*. [Jones-Rhoades et al., 2006] proposed a more stringent set of rules for non-conserved miRNA precursors and two of the new putative miRNAs do not pass those rules (Put-miR-2 and 3). Although Put-miR-1 fulfils the stringent rule, additional proof is required for a convincing miRNA classification, such as identifying the DCL family member that generates this sRNA. However, the lack of DCL mutant tomato or other *Solanaceous* species makes it impossible to generate such data at the moment. It will be more informative to identify target genes for cloned sRNAs, although this also requires a more complete genome sequence. Another approach is to generate a much larger sRNA sequence database by high-throughput sequencing and analyse the pattern of cloned sRNAs for a candidate miRNA locus. Only the mature miRNA and the miRNA* sequences should be found from a real miRNA locus. Our library is not large enough to confirm the putative miRNA loci since all three have been found only once.

The importance of this study is the identification of several sRNAs that are not present in *Arabidopsis*. Considering that only 4,018 sRNA were sequenced this result suggests that many more sRNAs specific to *Solanaceous* plants will be discovered

by high-throughput sequencing of tomato sRNA libraries as is indeed confirmed in Chapter 5. In addition, one of the putative miRNAs shows a significantly higher accumulation in fruit than in leaf, suggesting a specific role in fruit development/ripening and linking the sRNA pathway to an agronomically important tissue. This study lays the foundation for understanding the complexity of the sRNA population in tomato, attainable through the combination of high throughput sRNA sequencing and additional genome sequencing.

4.7 Discussion

The work described in this chapter involved the analysis of a small scale sRNA sequencing experiment which acted as a test case for the `miRCat` package described in the previous chapter. It enabled us to test and refine bioinformatics tools for both miRNA detection and target prediction before running on large scale, high-throughput datasets that we will discuss in the next chapter. It also led to the discovery of some interesting tomato specific sRNAs that showed differential expression patterns in fruit and leaf tissues.

Chapter 5

Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening

This chapter is an adapted and extended version of

Moxon, S., Runchun, J., Szittya, G., Schwach, F., Rusholme Pilcher, R.L., Moulton, V., Dalmay, T. (2008): Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening. *Genome Research.* In press

5.1 Summary

This chapter describes the analysis of 454 high-throughput sRNA datasets from tomato and can be regarded as a follow up study to the small scale sRNA cloning described in Chapter 4. All experimental work (Northern blots and 5' RACE analysis) was carried out by members of Dr. Tamas Dalmay's laboratory.

5.2 Background

To date, most plant miRNAs have been identified by the traditional Sanger sequencing method in *Arabidopsis*, rice and poplar, and comparison of miRNA sequences across plant families has shown that the majority of miRNAs are conserved [Axtell and Bartel, 2005]. However, some miRNAs appear to be species specific and Allen *et al*. [Allen et al., 2004] have suggested that these miRNAs have evolved recently ("young" miRNAs), in contrast to the conserved miRNAs ("old" miRNAs). Non-conserved miRNAs are often expressed at a lower level than conserved miRNAs, and this is one of the reasons why small-scale sequencing reveals mainly conserved miRNAs. As mentioned in Chapter 2, development of high-throughput pyrosequencing technology has allowed the discovery of several non-conserved or lowly expressed miRNAs through deep sequencing, e.g. in *Arabidopsis* and wheat [Rajagopalan et al., 2006, Fahlgren et al., 2007, Yao et al., 2007].

Since most plant developmental processes involve miRNA regulation

[Kidner and Martienssen, 2005], the discovery of non-conserved miRNAs suggests that plant species/families with specific developmental features may contain non-conserved miRNAs that are involved in the regulation of gene expression specific to those features. To investigate this hypothesis, we chose fleshy fruit formation and ripening as specific developmental features that are not characteristic to *Arabidopsis*, rice or poplar. Therefore if miRNAs are involved in these processes they should probably not be present in these species.

5.3 Materials and methods

5.3.1 Cloning of small RNAs, Northern-blot and 5'RACE analysis

Total RNA was extracted from tomato leaf, bud before flower blooming and different developmental stages of whole fruits. Small RNA between 19-24nt were cloned from leaf and fruit (mixture of different sizes between 1 and 15 mm) as described by Pilcher *et al*. [Pilcher et al., 2007]. Briefly, the sRNA fraction was purified and ligated to adaptors without de-phosphorylating and re-phosphorylating the sRNA. The RNA was converted to DNA by RT-PCR and the DNA was sequenced by 454 Life Sciences. Twenty micrograms of each total RNA sample was used for Northern blot analysis as described by Pall *et al*. [Pall et al., 2007]. 5'RACE analysis was carried out using poly(A) plus fraction and the GeneRacer kit (Invitrogen).

5.3.2 Bioinformatics analysis

Small RNA sequences were extracted from raw reads matching both the last seven nucleotides of the 5' adaptor and the first seven nucleotides of the 3' adaptor sequences. Sequences were then queried against ribosomal and transfer RNAs from Rfam [Griffiths-Jones et al., 2005], the *Arabidopsis* tRNA database `http://lowelab.ucsc.edu/GtRNAdb/Athal/` and rRNA sequences obtained from EMBL [Cochrane et al., 2006]. Any sRNAs having exact matches to these sequences were excluded from genomic mapping. Reads of 18-30nt were mapped to tomato BAC sequences (bacs.v175.seq) obtained from the SOL Genomics Network [Mueller et al., 2005] using exact matching. sRNAs were then analysed using `miRCat` `http://srna-tools.cmp.uea.ac.uk/mircat/`. Target predictions were performed based on methods described by Allen *et al*. [Allen et al., 2005] as implemented at `http://srna-tools.cmp.uea.ac.uk/targets/`

5.4 Results

5.4.1 Deep sequencing of tomato short RNAs

Two separate sRNA libraries were generated from mixed size (1-15mm) green tomato fruits of MicroTom, a miniature rapid-cycling cherry tomato variety [Meissner et al., 1997]. In addition, two sRNA libraries were prepared from tissue of young leaves of the same cultivar. The four libraries were sequenced by 454 Life Sciences using pyrosequencing technology that produced 721,874 reads yielding

402,197 and 168,570 sequences from fruits and leaves, respectively, with recognisable adaptor sequences (Table. 5.4.1). These reads represented around 225,000 and 102,000 unique sRNA sequences in fruits and leaves. In both tissues the 21nt and 22nt classes showed the highest degree of redundancy (Figure. 5.1 and Figure. 5.2), suggesting that sRNAs in these size classes are often produced from precursors from which clearly defined mature short sequences are excised. These sRNAs are often miRNAs and trans-acting siRNAs (ta-siRNAs) that are usually expressed at a high level [Vaucheret, 2006]. The 23 and 24nt classes were much less redundant (Figure. 5.1 and Figure. 5.2), indicating that they derive from loci that produce heterogeneous sRNA populations, such as those found associated with RNA-polymerase IV dependent pathways in *Arabidopsis* which produce heterochromatin-related siRNAs. To compare sequence redundancy levels in samples of different size, we normalised the larger fruit sample to the number of reads in the leaf sample by extracting 1000 random subsets of 159,886 reads from the fruit sample. Figure. 5.3 and 5.4 show size distributions of the leaf sample in comparison to the random average of the normalised fruit samples. The distribution of redundant sequences for different size classes was similar in fruits and leaves (Figure. 5.3). However, the size distribution of non-redundant sRNAs was slightly different in the two tissues (Figure. 5.4). The non-redundant leaf sRNA distribution showed a peak at 21nt, while there were more non-redundant fruit sRNAs of 22, 23 or 24 than 21nt. Assuming that the overall proportion of 24nt sRNA is related to the extent of transcriptional regulation, this observation suggests a more

extensive regulation of gene expression by sRNAs at transcriptional level in fruit than in leaf. This is probably because the longer sRNAs are often associated with DNA methylation and heterochromatin formation.

Table 5.1: Statistics of sRNAs sequences from tomato fruit and leaf

Fruit	**Reads**	**match BACv175**	**Unique reads**	**match BACv175**
Raw reads	537036			
Adaptors removed	402197	79099	224823	39001
rRNA/tRNA exact matches removed	391119	75353	222391	38064
Match known miRNAs	14536	3974	588	66
sRNAs mapping to predicted hairpins	3909	3909	409	409
Predicted hairpins sRNA abundance $\geqslant 3$	3766	3766	30	30
Leaf	**Reads**	**match BACv175**	**Unique reads**	**match BACv175**
Raw reads	184838			
Adaptors removed	168570	35419	102753	18180
rRNA/tRNA exact matches removed	159886	32535	100168	17032
Match known miRNAs	18352	5488	780	89
sRNAs mapping to predicted hairpins	4019	4019	235	235
Predicted hairpins sRNA abundance $\geqslant 3$	3766	3766	30	30

5.4.2 Known miRNAs

We searched for known miRNAs in our combined (fruit and leaf) tomato sRNA database and found 7,912 redundant sequences matching 20 known miRNA families (Table. B.1 and B.2 with precursor sequences shown in B.3). In addition we identified 25,436 sequences that were either shorter/longer or contained up to two mismatches to the same 20 and another 10 known miRNA families. Of these 30 families three had previously been thought to be specific to *Arabidopsis* (miR858; [Fahlgren et al., 2007]), algae (miR1151; [Molnár et al., 2007]) and moss (miR894; [Fattash et al., 2007]) and

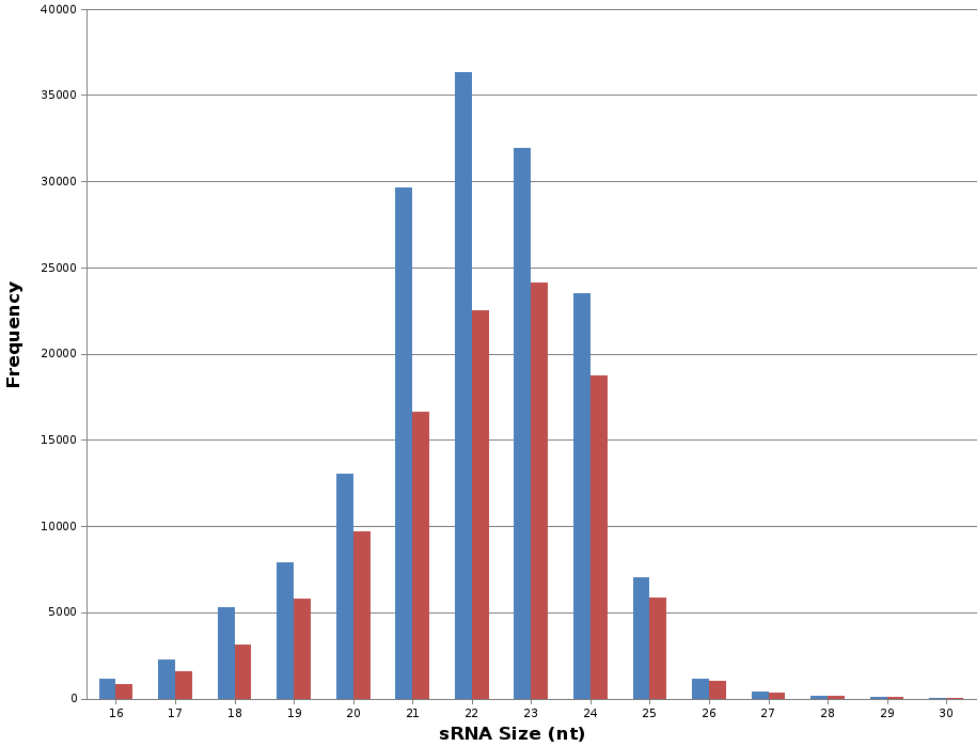

Figure 5.1: Histogram showing abundance/cloning frequency of redundant (blue) and non-redundant/distinct (red) sRNA reads from fruit samples.

were selected for testing by Northern blot. The algae specific miR1151, gave negative result and was probably an artifact. However, we were able to confirm the expression of miR858 and miR894 (Figure. 5.5). We also confirmed our previous observation [Pilcher et al., 2007] that miR482 (originally reported to be poplar-specific [Lu et al., 2005b]) is also expressed in tomato. These three examples show that miRNAs previously believed to be species or family specific can exist in several families. Data from more species is necessary to understand the evolution of these less conserved miRNAs.

We analysed the expression levels of 13 additional known miRNAs that were

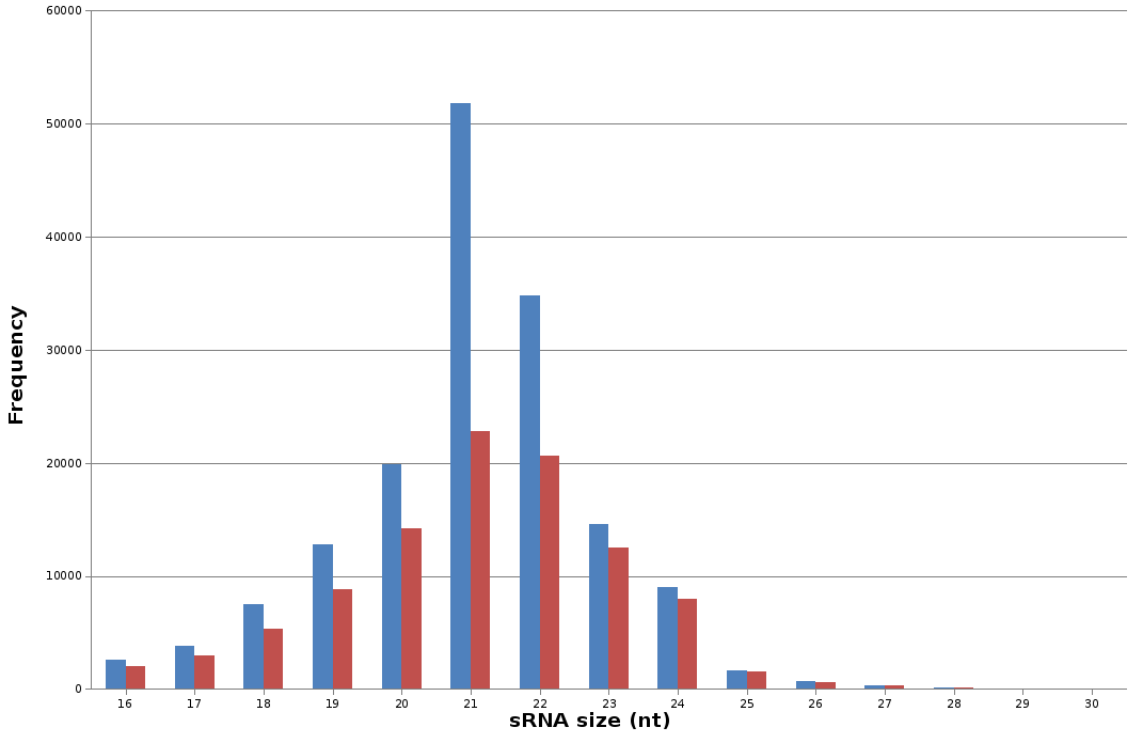

Figure 5.2: Histogram showing abundance/cloning frequency of redundant (blue) and non-redundant/distinct (red) sRNA reads from leaf samples.

present in our libraries and that had not been examined in our previous study [Pilcher et al., 2007] using Northern blot assays of samples from leaves, closed flower buds and four different stages of fruits (Figure. 5.5). All tested miRNAs, except for miR165/166, 403 and 472, showed differential expression patterns in these tissues. Several miRNAs (miR156/157, 164, 408, 858 and 894) were more abundant in leaves and closed flowers than in fruits. In contrast, miR169 was expressed at a higher level in all fruit stages than in closed flowers and it was almost undetectable in leaves. Intriguingly, two known miRNAs showed differential expression between different fruit stages. miR171 (and miR171*) was as highly expressed in very small fruits (1-3 mm)

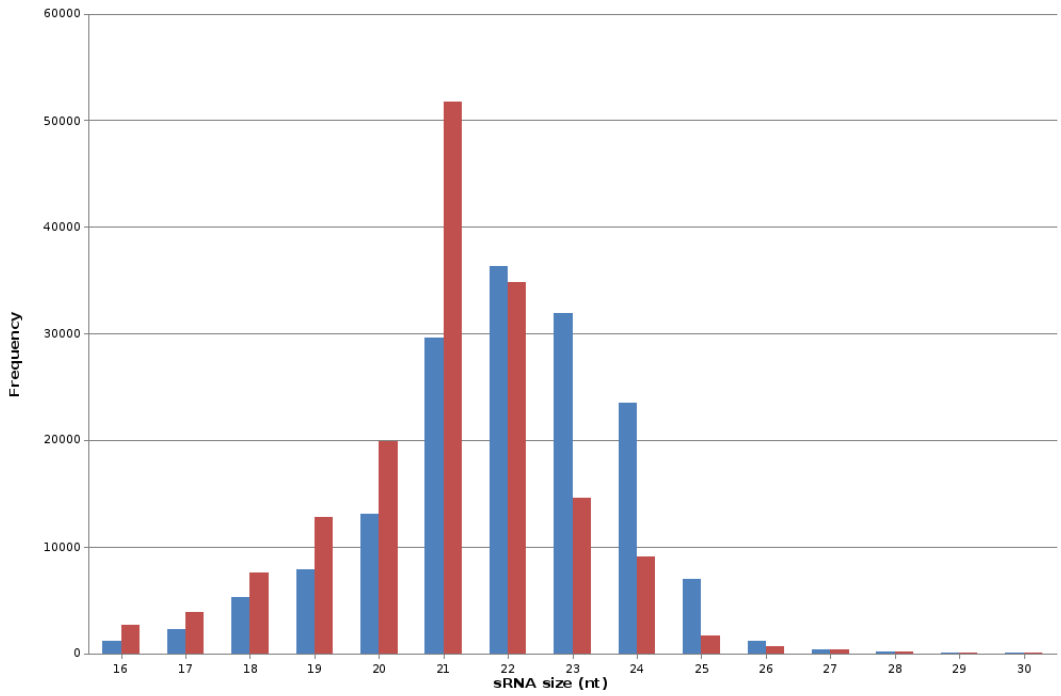

Figure 5.3: Histogram showing the normalised abundance/cloning frequency of redundant fruit (blue) and leaf (red) sRNA reads.

as in leaves and closed flowers, and it accumulated at a lower level in more mature fruits. Interestingly, miR390 had much higher accumulation in very small fruits than in leaves and closed flowers, and it accumulated at a very low level in more mature fruits. This suggests that miR390 has a specific role in early fruit formation.

Several target genes of known miRNAs have been validated in *Arabidopsis*, rice and poplar. However, it is not obvious which genes are targeted by these miRNAs in tomato because annotation of the partial genome sequence is not complete. In addition, Itaya *et al*. [Itaya et al., 2008] could only validate one out of three conserved miRNA target tomato genes (that miR172 targeted APETALA2). We used the tomato

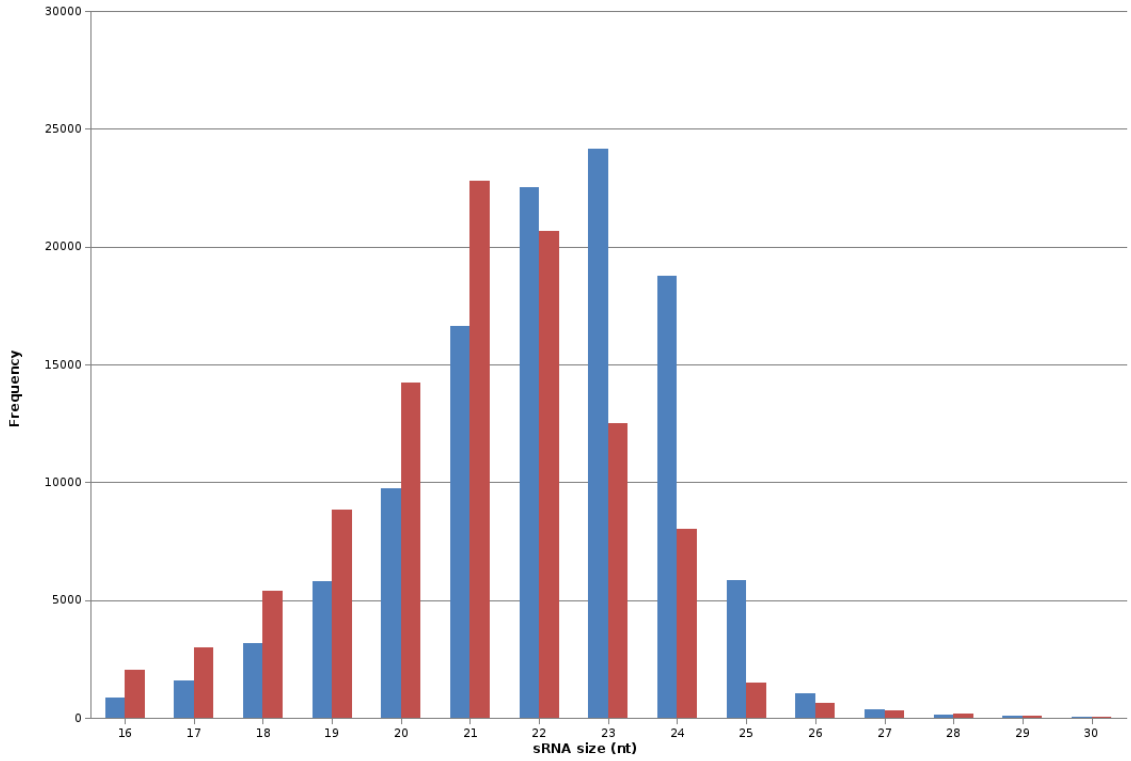

Figure 5.4: Histogram showing the normalised abundance/cloning frequency of non-redundant/distinct fruit (blue) and leaf (red) sRNA reads.

Unigene EST database [Mueller et al., 2005] to predict twelve targets that were all validated by 5' RACE assays (Figure. 5.5). Two targets are worth describing specifically; a MYB transcription factor that is targeted by miR858 (originally thought to be *Arabidopsis* specific [Fahlgren et al., 2007]), and Colorless non-ripening (CNR), a member of the Squamosa-promoter Binding Protein (SBP) family that was shown to be involved in fruit ripening [Manning et al., 2006]. CNR is targeted by miR156/157, which, for the first time, implies miRNA involvement in the maturation process of a commercially important fruit.

5.4.3 Novel miRNAs

As mentioned above, tomato genome sequencing is not yet complete, although many genomic BAC sequences are available [Mueller et al., 2005]. We used version BACv175 (unfinished) for our analysis that represents approximately 25% of the tomato genome. sRNA sequences that were not known miRNAs were mapped to BAC sequences. Secondary structures were predicted for each locus, and the ones that fulfilled the hairpin structure criteria described by Jones-Rhoades *et al*. [Jones-Rhoades et al., 2006] were selected as candidate miRNAs. This analysis resulted in 219 (165 unique) candidates, 87 out of which also had a predicted target in at least one tomato EST sequence. We also looked for sequenced miRNA* sequences. However, most of the potential mature miRNAs were sequenced less than five times and so no miRNA* sequences were found (average frequency of miRNA* is around 10% of the frequency of mature miRNA [Rajagopalan et al., 2006]). However, one sequence was found 19 times and a potential miRNA* was sequenced 9 times in our combined datasets. According to the criteria published by Rajagopalan *et al*. [Rajagopalan et al., 2006], this sequence, is the first novel bona fide miRNA (Sly-miR-Z) identified in tomato. The other miRNA candidates were further tested by northern blot (miRNA) and 5' RACE assay (target). Northern blot analysis was carried out for 92 candidates with hairpin structure but without sequenced miRNA*. 51 were detected as discrete bands around 21nt and several showed differential expression in different tissues (Figure. 5.7). Three sRNAs showed very strong leaf specificity,

one sRNA accumulated at higher level in closed flowers and all stages of fruits than in leaves, and seven sRNAs showed increasing level of expression during fruit development, suggesting the possibility that several genes involved in fruit development are regulated by short RNAs.

Many miRNA candidates had predicted targets. We carried out 5' RACE assay for 65 predicted targets, but most of them (62) could not be validated. Therefore, these sRNAs cannot be classified as miRNAs because, although they were sequenced, produced from stable hairpins, and accumulated as 21nt RNA, no miRNA* was sequenced, no target cleavage was shown and their accumulation in a dcl1 mutant could not be studied due to the lack of such a mutant in tomato. However, we strongly suspect that several of these potential miRNAs are bona fide miRNAs, although at least one of the above-mentioned criteria would have to be shown to hold before they could be classified as miRNAs.

In addition to Sly-miR-Z, we found three new tomato miRNAs (secondary structures are shown in Figure. B.4). Although no miRNA* sequences were found for these three miRNAs, their predicted target genes were validated by 5'RACE analysis (Figure. 5.8B). The target genes of Sly-miR-X are the splice variants of constitutive triple response 4 (LeCTR4sv1 and LeCTR4sv2), which is a member of the CTR family that are key negative regulators of ethylene responses [Adams-Phillips et al., 2004]. Both Sly-miR-Y and Sly-miR-W target unknown protein expressing mRNAs (ESTs SGN-U326398 and SGN-U322371, respectively) that do not show any homology to

annotated genes in the EMBL sequence database. Accumulation of the new miRNAs was analysed by Northern blot, and Sly-miR-Y and Sly-miR-Z showed significantly stronger expression in fruit than in leaf or flower bud (Figure. 5.8A). In fact, these two miRNAs accumulated at a higher level in more mature fruit than in very young fruit. Sly-miR-X produced a consistently weak signal and it was necessary to use an LNA (locked nucleic acid) probe to reveal stronger accumulation in more mature fruit. Sly-miR-W is expressed at a similar level in all analysed tissues (Figure. 5.8A). The complete genome sequences of *Arabidopsis*, rice and poplar were interrogated for miRNA genes homologous to the four new tomato miRNAs but no perfect matches were found. A few loci were found with two mismatches but none of these exhibited a hairpin structure with their flanking regions (data not shown). We therefore concluded that the new tomato miRNAs were not conserved in these species.

5.4.4 Other tomato specific sRNAs

Most sequenced *Arabidopsis* sRNAs are 24nt and derived from transposons and other repeats [Rajagopalan et al., 2006, Fahlgren et al., 2007, Mosher et al., 2008]. We found that the 24nt class of sRNAs was also generally abundant in tomato, especially in fruit. However, the abundancy of sRNA sequences from one particular class of transposons was exceptional. 9 280 sequences were derived from type III foldback transposon Tomato Anionic Peroxidase Inverted Repeat [Hong and Tucker, 1998]. TAPIRs are approximately 280nt long inverted repeats, often located adjacent to genes [Mao et al., 2001]. This element has a high copy number; we found 468

copies in the available genome sequence (25% of the genome). Although sRNAs are usually well dispersed over transposon sequences, we found several sRNAs that mapped to TAPIRs derived from specific locations of the inverted repeat (Figure. 5.9A). In fact, most TAPIR loci produced sRNAs predominantly from two regions that were opposite to one another on the two stems of the hairpin structure (reminiscent of potential miRNA/miRNA* pairs). However, the most abundant sequences from the two regions did not form a precise miRNA/miRNA* duplex (Figure. 5.9A). We tried to compare the distribution of sRNAs derived from TAPIR to the accumulation pattern of sRNAs derived from a type III foldback transposons in *Arabidopsis* [Adé and Belzile, 1999], but the published *Arabidopsis* sRNA databases contained almost no sRNA sequences derived from hairpin elements [Rajagopalan et al., 2006, Fahlgren et al., 2007, Mosher et al., 2008, Qi et al., 2006]. Next we searched for hairpins in the *Arabidopsis* genome that are longer than 200nt and produce sRNAs. The sRNA pattern of these loci was different from TAPIRs; sRNAs were scattered across the whole hairpin and were absent in dcl-4 plants. Next we compared the positions of the most abundant sRNAs derived from different TAPIR loci. The most abundant sRNA sequence was different in some TAPIR elements, although some loci produced the same major sRNA. Northern blot analysis of the two most abundant TAPIR sRNAs gave a slightly different expression pattern in spite the fact that they were shifted only by two nucleotides. TAPIR1 accumulated at a slightly higher level in leaf than fruit, and TAPIR2 was more abundant in fruit than in leaf

(Figure. 5.9B).

The other new class of sRNAs, which was not found in libraries from other species, derived from endogenous pararetroviral (EPRV) sequences. Several DNA viruses were found integrated into the host genome, some of which can cause infection and some not [Harper et al., 2002]. An EPRV was recently described in tomato that was proposed to be controlled by RNA silencing through sRNAs [Staginnus et al., 2007]. Several sRNA sequences matched an integrated EPRV sequence but, surprisingly, they were not randomly distributed. One particular sequence (EPRV1) was found with a very high frequency in all four libraries in addition to a few less abundant hot-spots. Although Northern blot analysis confirmed the accumulation of EPRV1 and two other less abundant EPRV specific siRNAs, the very high frequency of EPRV1 was not reflected by the Northern blot (Figure. 5.10). In fact, EPRV3 was easier to detect than EPRV1. Expression analysis of EPRV specific siRNAs in different cultivars and wild species showed that their expression varies in different accessions although integrated copies were detected in all of them [Staginnus et al., 2007].

5.5 Conclusions

5.5.1 Conserved miRNAs in tomato

We generated sRNA libraries from fruit and leaf of tomato plants, and most conserved miRNA families were found in at least one of our sRNA libraries. Several conserved miRNAs showed differential expression in leaf, flowering bud and fruits at different

stages that could provide information about their function. Sly-miR-169 was preferentially accumulated in flower buds and fruits and was hardly detectable in leaves (Figure. 5.5). The only known target of this miRNA is a transcription factor of the CCAAT-binding family, HAP2 and this protein was shown to have an important role during nodule development in *Medicago truncatula* [Combier et al., 2006]. HAP2 is also required for pollen tube guidance and fertilisation [von Besser et al., 2006] and affects flowering time [Wenkel et al., 2006] in *Arabidopsis*. It remains to be seen whether HAP2 also plays a role in fruit development or if Sly-miR-169 can target other genes in tomato. Another miRNA highly expressed in fruits is Sly-miR-390. In fact, its expression sharply peaks in very young fruits (1-3 mm; Figure. 5.5) suggesting that it plays a role in early fruit development. Sly-miR-390 is also expressed at a lower level in leaves, which is in line with its known function in *Arabidopsis* where it controls leaf morphology through targeting TAS3 [Adenot et al., 2006]. Cleaved TAS3 gives rise to ta-siRNAs targeting mRNAs in the AUXIN RESPONSE FACTOR (ARFs) family [Allen et al., 2005] and it will be interesting to see whether TAS3 or other TAS gene derived ta-siRNAs are involved in early fruit development (a TAS3 homologue is present in the tomato genome). In contrast to Sly-miR-390 and 169, Sly-miR-894 was hardly detectable in fruits but accumulated at a high level in flower buds and leaves (Figure. 5.5). Interestingly, this miRNA was only found in moss previously but no target gene was identified [Fattash et al., 2007]. Sly-miR-408 was also absent

in fruits and it does not have any validated targets, although it was predicted to target plantacyanin genes [Sunkar and Zhu, 2004]. These examples show that target identification and validation of conserved miRNAs in different species is still important because miRNAs may have additional function to the regulation of conserved target genes. Axtell *et al*. [Axtell et al., 2007] demonstrated that different sRNAs in different species can have similar functions. New targets for conserved miRNAs may be found in tomato after genome sequencing is completed but at this time we could only use the available Unigene sequences downloaded from the SOL Genomics Network [Mueller et al., 2005]. Twelve target genes were validated for nine conserved miRNAs (Figure. 5.6). One of these miRNAs, Sly-miR-858, is not present in the poplar and rice genome [Rajagopalan et al., 2006] but was found in *Arabidopsis* where it targets the MYB12 transcription factor. We identified two MYB12-like genes in tomato and both of them showed the same level of similarity to the *Arabidopsis* gene. One of them validated better (SGN-U320618) than the other (SGN-U322556) which showed a scattered cleavage pattern around the canonical cleavage site. This pattern was very similar to the cleavage sites of the Sly-miR-172 targeted APETALA2-like tomato gene (SGN-U314858; Figure. 5.6), raising the possibility that both SGN-U322556 and SGN-U314858 (AP2) are primarily suppressed at the translational level similar to the miR-172 mediated APETALA2 regulation in *Arabidopsis* [Chen, 2004]. Alternatively, mRNAs showing several cleavage sites around the canonical sites are not targeted by

miRNAs. Another transcription factor family regulated by miRNAs is the Squamosa-promoter Binding Protein (SBP) family. The tomato members of this family are called SBP-like proteins (SPL). We validated three SPL genes targeted by Sly-miR-156/157 and one of them is CNR, a key gene in fruit ripening [Manning et al., 2006].

5.5.2 Classification of non-conserved miRNAs

Recently a number of publications have reported high-throughput sequencing of sRNAs from *Arabidopsis* [Lu et al., 2005a, Lu et al., 2006, Rajagopalan et al., 2006, Fahlgren et al., 2006, Mosher et al., 2008] and other plant species [Yao et al., 2007, Axtell et al., 2007, Molnár et al., 2007, Barakat et al., 2007a, Barakat et al., 2007b, Morin et al., 2008]. The common theme emerging from these reports is that the sRNA content of plants is very complex and, although a subset of sRNAs is conserved across different families, a number of sRNAs are specific to each species or family. The most conserved class of sRNAs is the miRNA class but even these are not all conserved. These observations led to a change in the minimum criteria for classifying an sRNA as a miRNA initially set up by Ambros *et al*. [Ambros et al., 2003]. Even so, the fact that many of the loci that express sRNA can be folded into a stem loop structure prompted Rhoades *et al*. [Jones-Rhoades et al., 2006] to introduce new criteria to avoid flooding miRBase with sequences that are not miRNAs. In particular, the conservation criterion was replaced with proof of biogenesis (demonstration of DCL1 dependency or cloning of perfect miRNA* sequences) or functional data (target validation by 5' RACE). However, this criterion was not verified in several recent studies partly because a dcl1

mutant was not available for species other than *Arabidopsis*. Since sequence complementarity between miRNAs and their target genes are very high in plants, target validation has been increasingly overlooked, and a number of recent studies have considered target prediction as sufficient functional data. This is probably due to the fact that, at least initially, all predicted targets that were experimentally tested proved to be real targets. However, most validated targets are recognised by conserved miRNAs and the predicted targets of most non-conserved miRNAs have never been tested experimentally. Here we show that most predicted targets of putative non-conserved miRNAs could not be validated experimentally, in contrast to the high validation rate of targets of conserved miRNAs. There are several possible explanations for negative 5' RACE results, such as the target genes are not expressed in the same cells as the putative miRNAs or the cleavage product is not stable enough. However, it is more likely that many of the putative miRNAs are false positive predictions and not true miRNAs. They are expressed and could derive from hairpin structure precursors but it is now clear that these criteria hold for many thousands of loci in plant genomes, and it does not necessarily mean that they do derive from single-stranded stem-loop structures. In the absence of biogenesis data, it has to be shown that the potential miRNAs mediate cleavage of mRNAs in order to classify them as miRNAs [Jones-Rhoades et al., 2006]. Our observation suggests that recently published non-conserved miRNAs predicted by high-throughput sequencing projects have to be considered cautiously because many of them are likely to be siRNAs and not miRNAs.

We validated cleavage of three novel targets mediated by new non-conserved tomato miRNAs (Figure. 5.8B), although one of them produced a less precise cleavage pattern. This pattern is similar to the cleavage pattern of miR172 and miR858 targets and so this target may be primarily suppressed at translational level. A fourth new miRNA was validated by cloning of the perfect miRNA* sequence. One of the novel targets is a member of the CTR gene family that suppresses ethylene response and is involved in fruit ripening [Adams-Phillips et al., 2004]. This result, together with the regulation of CNR by Sly-miR-156/157, opens a new avenue in the field of gene expression regulation during fruit development and ripening.

5.5.3 Can some miRNA genes derive from transposons?

Transposon specific sRNAs are usually abundant in sRNA libraries but sRNAs derived from type III foldback transposon TAPIR sequences [Hong and Tucker, 1998] were exceptionally highly represented in the two tomato sRNA libraries. TAPIR elements are flanked by nine nucleotide target site duplications and they are mobile [Mao et al., 2001]. However, sRNAs are not well dispersed over TAPIR elements like on other transposons. Instead, they map to specific positions that would be on opposite arms of a hairpin structure if the TAPIR element is expressed (Figure. 5.9). The most abundant sRNA species are similar to miRNA/miRNA* pairs but they do not precisely pair with each other. Moreover, there are less frequently sequenced sRNAs around the most abundant species, although the pattern of sRNAs is likely to be less complex than it is shown on Figure. 5.9. Some sRNAs may not be

produced from the locus shown and but are instead produced from another TAPIR locus. Due to the high degree of sequence similarity between TAPIR elements, the sRNA will map to both loci resulting in a complex distribution of sRNAs on the hairpin. The sRNA pattern of TAPIRs is reminiscent of Ath-miR-822 and 839 two DCL4 processed non-conserved miRNAs [Rajagopalan et al., 2006]. However, those miRNAs are single locus genes in *Arabidopsis* and we found 468 copies of TAPIR in the available 25% of the tomato genome. We tried to predict target genes for TAPIR derived sRNAs but our analysis only found other TAPIR elements (data not shown). It is tempting to propose that TAPIRs are potential progenitors of miRNA genes and that if a TAPIR derived sRNA acquires a target gene and this regulation is beneficial for the plant that it could be fixed. Due to mutations, the miRNA producing TAPIR element could lose mobility and eventually become a proper miRNA gene. It was shown recently that several human miRNAs derive from transposable elements [Piriyapongsa and Jordan, 2007, Piriyapongsa et al., 2007] that support our prediction. According to the current model, miRNA genes are derived from target genes through duplication, inversion and mutations, and this is well supported by the fact that non-conserved miRNA genes often have low copy number and show extensive complementarity to the target gene beyond the mature miRNA sequence [Allen et al., 2004]. Transposons could be an alternative source of miRNA genes, especially in plant with large genomes and high copy number of foldback transposons.

5.6 Discussion

In this chapter we have presented results obtained using the tools described in Chapter 3 including `miRCat` and the target prediction software. We have discovered and experimentally validated the first tomato specific miRNAs and found two miRNAs that are potentially involved in fruit ripening as they regulate genes which are known to be important in fruit development. We have also discovered that TAPIR elements show miRNA-like sRNA accumulations and could act as a progenitor of new miRNAs in tomato. It is highly likely that a number of further novel miRNAs are present in the 454 datasets but due to the lack of a complete genome sequence they can not currently be detected computationally.

An important discovery in this work made from a bioinformatics perspective is that plant miRNA target prediction is not as accurate as generally assumed. Whilst target prediction algorithms may work in *Arabidopsis thaliana* where they have been thoroughly tested, the biological mechanisms in other plant species may differ implying that they are not so accurate. In this study many targets were predicted using the target prediction method outlined in Chapter 3 but most targets could not be validated experimentally suggesting a high degree of false-positive predictions. The alternative explanation for this could be that the targets are not cleaved but instead the mRNAs are translationally repressed (as is common in animals). A recent publication by Brodersen *et al*. [Brodersen et al., 2008] seems to confirm that this is true for several *Arabidopsis* mRNAs and perhaps this level of regulation is more common in tomato.

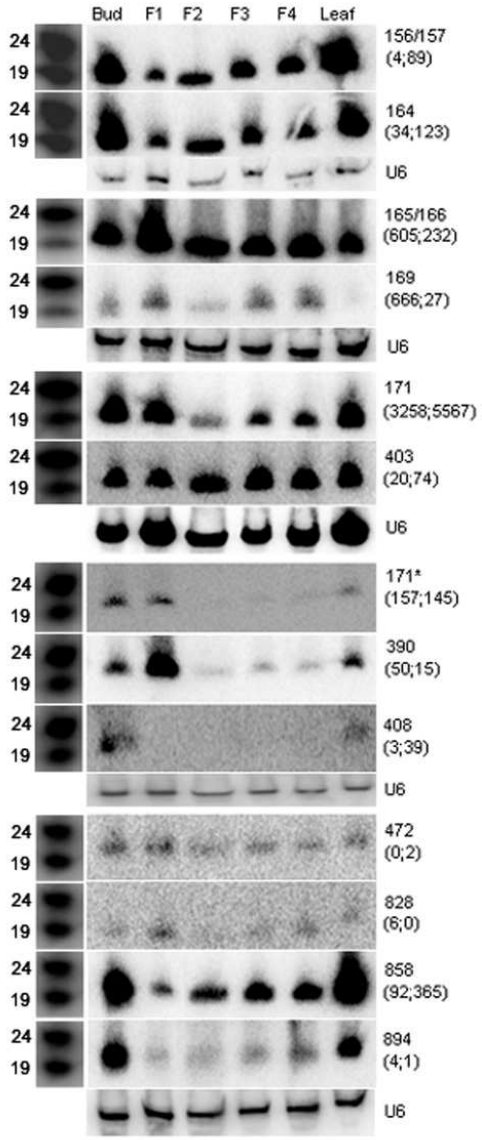

Figure 1

Figure 5.5: Expression of conserved tomato miRNAs: Total RNA from different tissues was extracted, separated and transferred to membranes. The membranes were hybridised to miRNA specific probes or a U6 specific probe (shown on the right) to demonstrate equal loading. Membranes were stripped and re-probed, equal loading is shown once for each membranes. Numbers between brackets indicate the number of sequences found in the fruit (left) and leaf (right) libraries for each miRNA. Different size fruits were used for RNA extraction; F1: 1-3mm, F2: 5-7 mm, F3: 7-11 mm and F4: 11-14 mm.

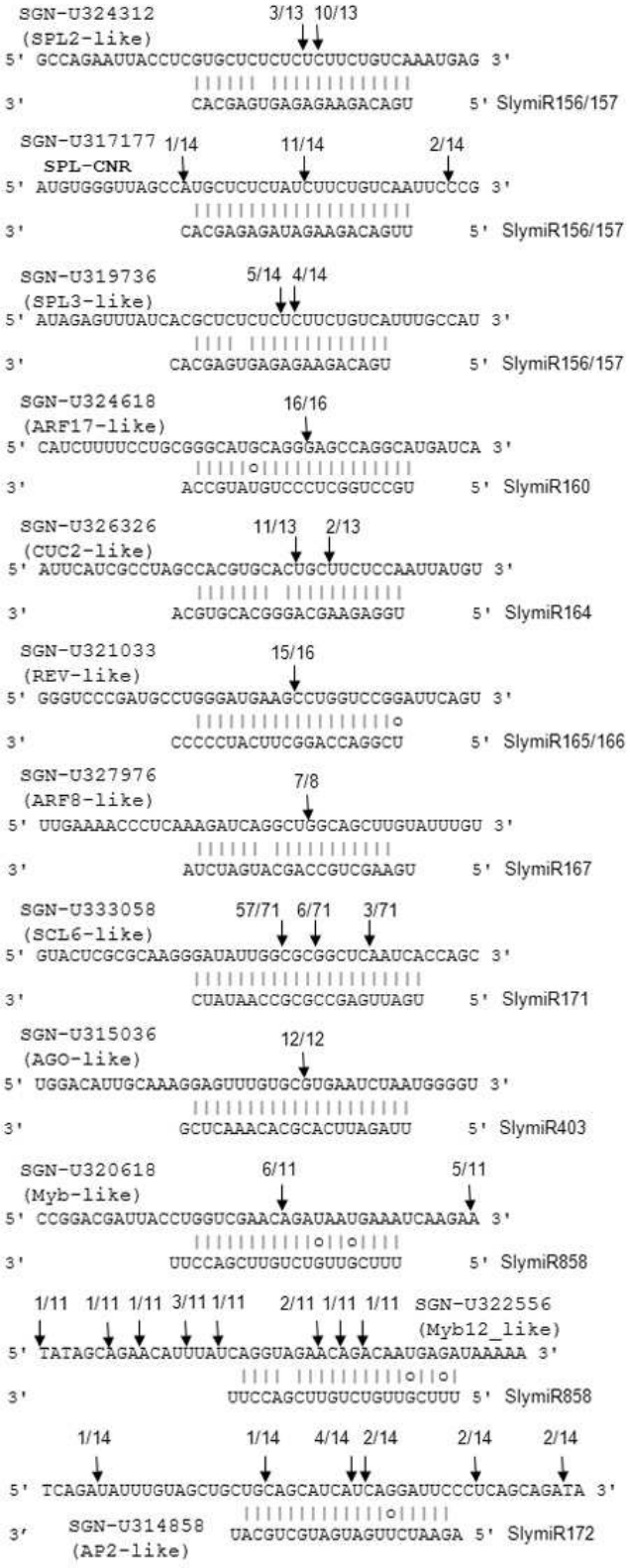

Figure 2

Figure 5.6: Target validation of conserved tomato miRNAs: 5'RACE analysis was carried out for each predicted target gene. Arrows show the 5' ends of cleavage products. Cleavage sites outside of the displayed sequence are not shown. Target EST sequences are shown on top of the miRNA sequences.

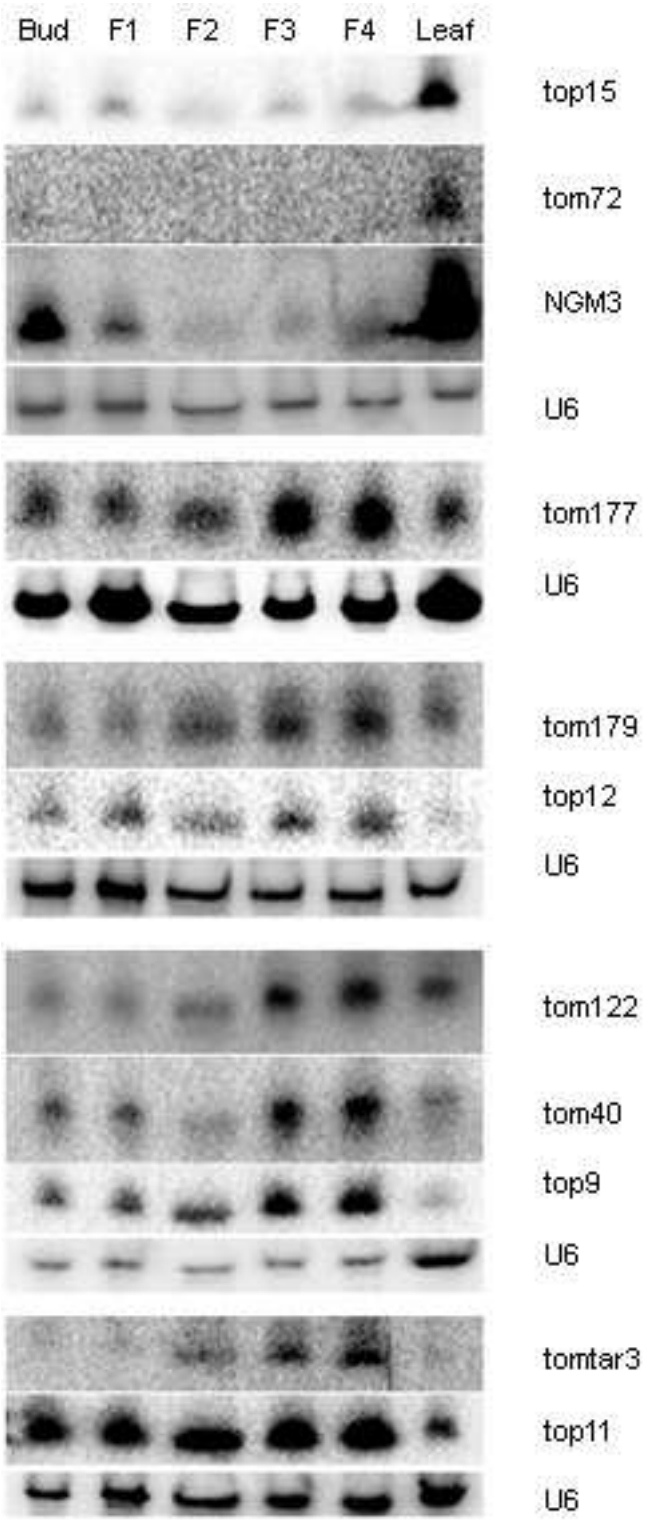

Figure 5.7: Differentially expressed tomato short RNAs: Probes specific to potential miRNAs (tom72, NGM3, tom177, tom179, tom122, tom40 and tomtar3) or short RNAs that could not be mapped to the available genome sequence but cloned many times (top15, top12, top9 and top11) were hybridised to the same membranes shown on Figure 5.5. U6 specific probe shows equal loading. Different size fruits were used for RNA extraction; F1: 1-3mm, F2: 5-7 mm, F3: 7-11 mm and F4: 11-14 mm.

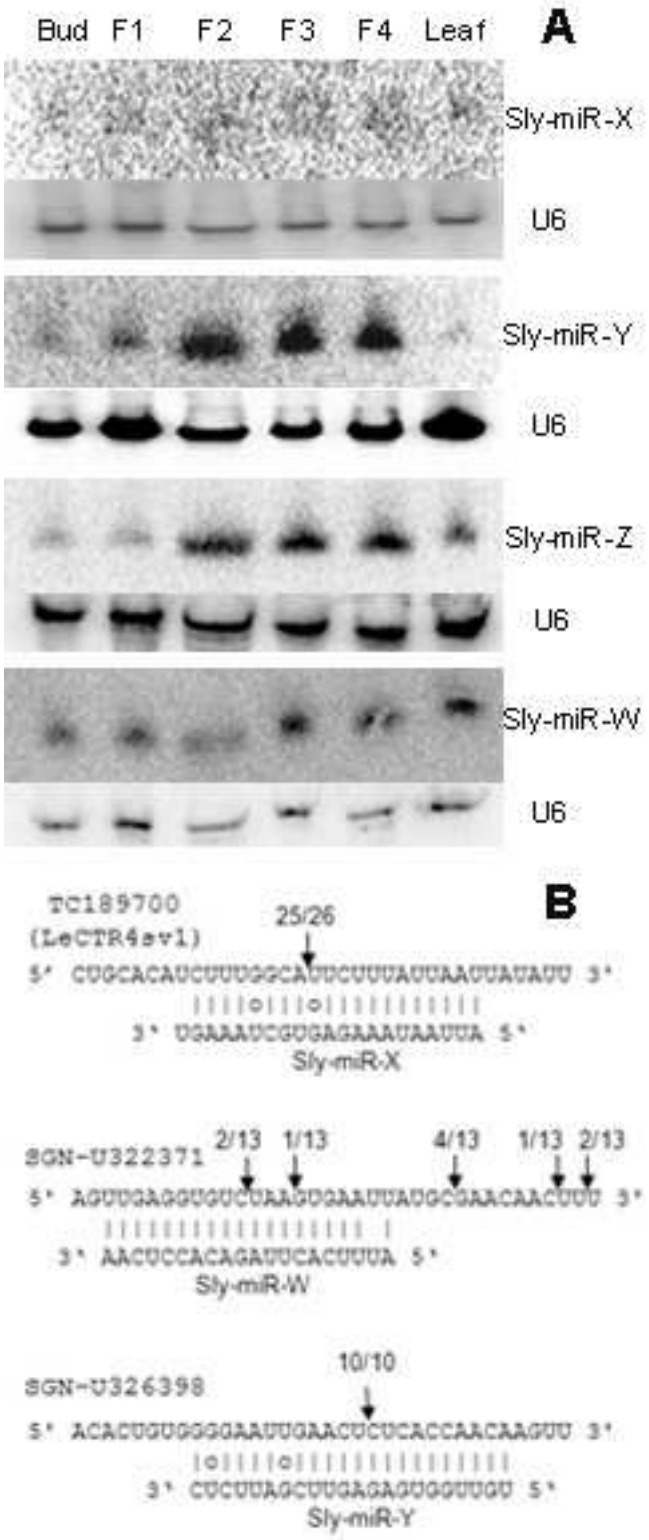

Figure 5.8: Expression and target validation of new non-conserved tomato miRNAs: Northern blot analysis of new miRNAs A) showed that Sly-miR-Y and Z accumulate preferentially in the fruit. U6 probe was used to show equal loading. Different size fruits were analysed; F1: 1-3mm, F2: 5-7 mm, F3: 7-11 mm and F4: 11-14 mm. B) shows the result of target validation for three new miRNAs. Arrows show the 5' ends of cleavage products mapped inside the displayed sequence. Target EST sequences are shown on top of the miRNA sequences.

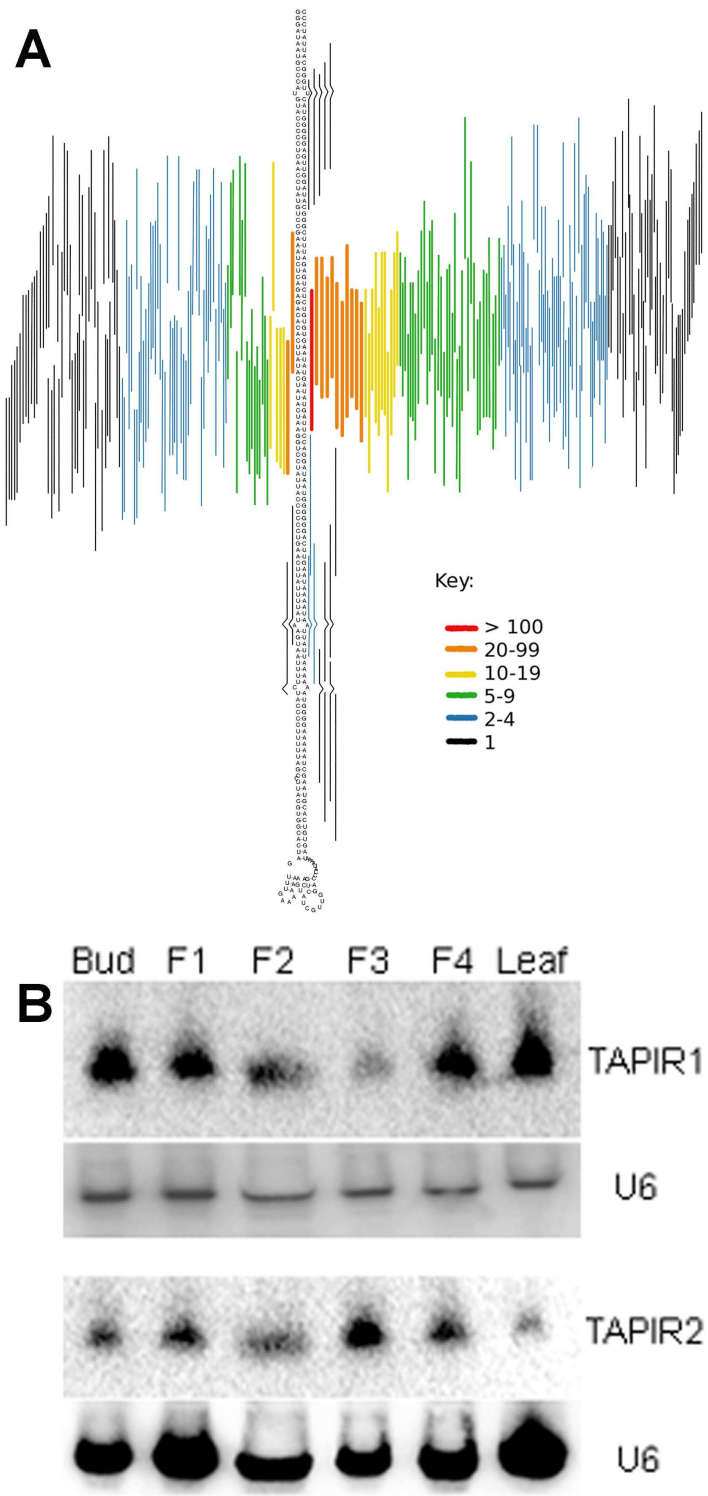

Figure 5.9: TAPIR derived sRNAs: A) Predicted secondary structure of one particular TAPIR element with lines representing the sRNA sequences mapping to the two arms of the hairpin. The colour of the lines specifies the abundancy of the sequences in the library. B) Northern blot shows the accumulation of the two most abundant, overlapping sRNAs from TAPIR elements. Membranes were stripped and re-probed for U6 to show equal loading. F1: 1-3mm, F2: 5-7 mm, F3: 7-11 mm and F4: 11-14 mm.

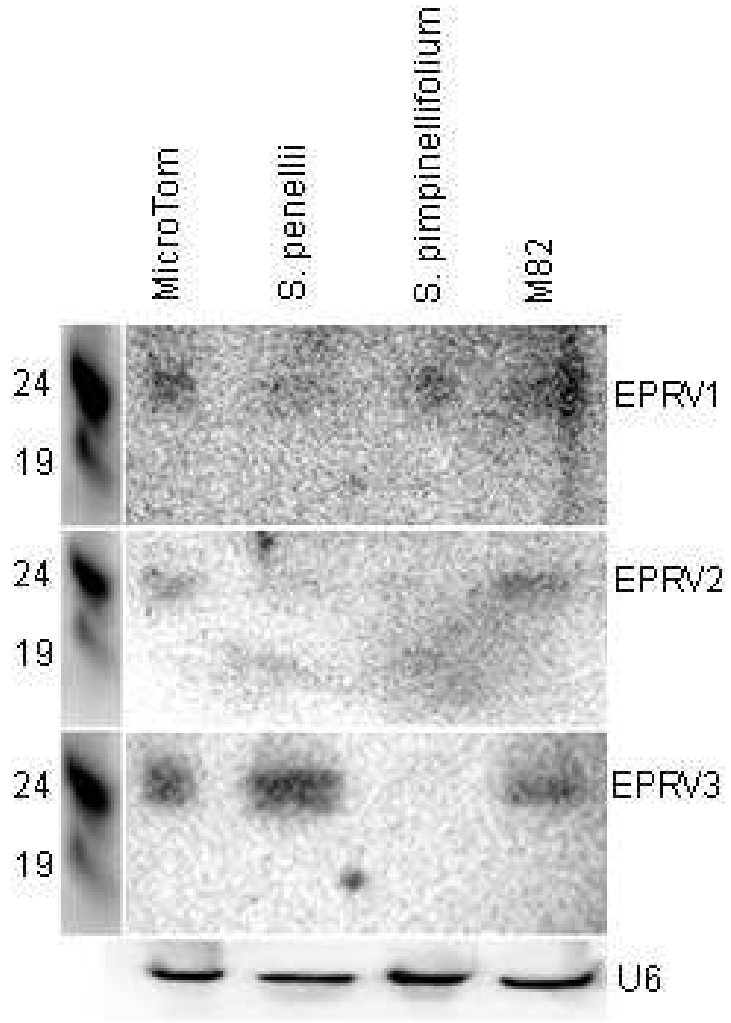

Figure 5.10: Expression of endogenous pararetrovirus specific sRNAs: Northern blot analysis of EPRV specific sRNAs shows the accumulation of mainly 24nt RNA species in leaves of four tomato accessions (MicroTom, *S. penellii*, *S. pimpinellifolium* and M82). 19 and 24nt RNA oligonucleotides were used as size markers (left) and a U6 probe shows equal loading.

Chapter 6

Identification of novel miRNAs in unsequenced genomes

6.1 Summary

All published miRNA identification algorithms, including those presented so far in this thesis, rely on a genome sequence in order to detect and classify miRNA hairpin precursors. Due to the comparatively small number of fully sequenced eukaryotic genomes and the great interest in the role of miRNAs in a number of different biological processes, it would be beneficial to devise new methods which do not require genome information. In this chapter we shall see that this type of “blind” approach to miRNA prediction is possible with the use of sRNA deep-sequencing data. In particular, we describe a new method that uses a support vector machine (SVM) classification approach to find putative miRNA/miRNA* pairs in sRNA deep sequence data and demonstrate its application to *Arabidopsis thaliana*, *Oryza sativa* and *Solanum lycopersicon* datasets.

6.2 Background

As explained in Chapter 2 mature miRNA sequences are derived from a longer hairpin-like structure (see Figure. 2.1). The mature miRNA along with a complementary miRNA* sequence are excised from the primary transcript by Dicer, and the miRNA is then incorporated into the RNA Induced Silencing Complex (RISC), after which the miRNA* is usually degraded. Deep-sequencing technology has allowed increasingly large numbers of sRNAs to be sequenced from a single sample so that miRNA* sequences are being routinely found in such datasets along with the more abundant miRNA sequences.

Since it has been previously impossible to predict miRNAs without using genome sequence information to predict secondary structures of putative precursors, miRNA research has generally been focused on organisms with a fully sequenced genome. Thus, for those researchers working with other, non-model, organisms there has been no way to predict novel miRNAs other than to map sRNAs of interest to related sequenced genomes, ESTs or other sequences deposited in EMBL and Genbank (e.g. [Sunkar and Jagadeeswaran, 2008]). However, as both miRNA and miRNA* are often present in high-throughput sRNA sets we can attempt to use these to find miRNAs without requiring a genome sequence.

Here we develop a support vector machine (SVM) approach to classify miRNAs based on miRNA/miRNA* pairs. SVMs are machine learning algorithms that attempt to use features from true and false examples in order to build a model which can then

be applied to classify new, unknown data without having to rely on fixed, rule based methods or expert knowledge. The SVM addresses the problem of learning to discriminate between positive and negative members of a given class by mapping all data points into an n-dimensional space (where n is the number of features selected) and then attempting to find a plane that separates the positive from the negative examples (the separating hyperplane) in the training set. Many potential separating hyperplanes may exist but the SVM is able to determine maximum-margin hyperplane which maintains a maximum margin from any point in the training set. Selecting the maximum-margin hyperplane maximises the SVMs ability to correctly classify previously unseen examples [Vapnik, 1998, Noble, 2006].

SVMs have been successfully applied in a variety of diverse fields including bioinformatics, where amongst other applications they have been used to classify both miRNA hairpins [Loong and Mishra, 2007a, Xue et al., 2005] and miRNA targets [Wang and Naqa, 2008, Kim et al., 2006].

6.3 Methods

Forty different *Arabidopsis thaliana* miRNA/miRNA* pairs were extracted from a Solexa sRNA dataset and used as a positive training set. Two sets of 100 different sRNAs with a read count of one were extracted from the Solexa dataset and shuffled using the shuffle program packaged with `SQUID` [SQUID, 2002]. This provided a

randomised set of sRNAs which retain the dinucleotide frequency of the original sequences. Preserving the dinucleotide frequency of the randomised RNA sequences is important since the secondary structure of a given RNA sequence is known to depend on dinucleotide base stacking energies as well as base pairing interactions [Workman and Krogh, 1999, Clote et al., 2005].

miRNAs were then searched against their complementary miRNA* sequences using FASTA [Pearson and Lipman, 1988] with features of the miRNA/miRNA* pair being encoded into a format readable by the `LIBSVM` package [Chang and Lin, 2001] to provide a true-positive training set. The two sets of randomised sRNAs were then searched against one other and all resulting pairs were encoded to provide the true-negative training set. Thirty-nine different features were selected in order to train the SVM which are described in detail in Appendix C.

The true positive and true negative examples were then combined and `LIBSVM` [Chang and Lin, 2001] was used to train the SVM and build the model. `easy.py` (included in the `LIBSVM` package) was then used to determine the optimal training parameters. The training data was normalised or “scaled” using `svm-scale`, a process that scales all feature values to lie in a range between -1 and 1, that is required for the training phase.

Five fold cross-validation was used for the training/testing phase in order to estimate the SVMs accuracy. In other words, the data was divided into 5 sub-samples and of these a single sub-sample was retained as the validation data for testing the model,

and the remaining 4 sub-samples were used as training data. The cross-validation process was then repeated 5 times, with each of the 5 sub-samples used exactly once as the validation data. The 5 results were then averaged to produce the accuracy estimation of the SVM predictions. On this test data 100% accuracy was obtained.

6.4 Results

The method was tested using several combined *Arabidopsis thaliana* 454 datasets taken from [Fahlgren et al., 2007, Rajagopalan et al., 2006, Mosher et al., 2008, Qi et al., 2006]. As the method is very computationally intensive, and we know from previous experience that miRNAs tend to be over-represented in high-throughput sequence data, we only tested a subset of the most abundant sRNA sequences as potential miRNAs. It is also known that miRNA* sequences tend to be cloned at a much lower frequency than miRNAs (around a 1:10 ratio of miRNA* to mature miRNA) [Rajagopalan et al., 2006] implying that if a miRNA is found at a low abundance then it is unlikely that its miRNA* will be present in the dataset. Hence all sequences with an abundance of 20 or more (total of 1922 non-redundant sequences) were used as input. These sequences were first searched against tRNAs, rRNAs and other non-coding RNAs (with the exception of miRNA families) from Rfam [Griffiths-Jones et al., 2005], and tRNAs from the Arabidopsis tRNA database [Lowe, 2004], and all sequences with two or fewer mismatches to any of these datasets were filtered out, leaving a total of 1705 unique input sequences. These 1705 sRNAs were then further filtered by size

to select for sRNAs in the range of 20-22nt (a total of 612 sequences) which is typical of most miRNAs (Figure. C.1).

The filtered input set of 612 sequences were searched against all sequences with a read count of five or more from the 454 high-throughput set (5136 non-redundant reads). Features were then extracted for each potential miRNA/miRNA* pair and the output from this stage was then scaled with `svm-scale` using the same parameters as used for the training set. The scaled feature set was classified with `svm-predict` (part of the `LIBSVM` package) using the model obtained from the training data. Each of the potential miRNA/miRNA* pairs were classified as either positive or negative by the SVM and assigned a p-value.

Of the 612 input sequences, 451 were classified as miRNA/miRNA* pairs with the default p-value of 0.5 or more. Each of the predicted miRNA/miRNA* pairs were then mapped to the *Arabidopsis* genome and those that appeared within 1000nt of one other and had the same orientation on the genome (i.e. were produced from the same genomic locus) were counted as miRNAs. Those pairs that came from different loci or different chromosomes were classed as false positives. In this way, 137 predictions were classed as miRNAs and the remaining 314 were classed as false positives as the predicted miRNA/miRNA* pairs could not be mapped to the same genomic locus.

Next a p-value filter was applied in order to try to reduce the number of false positive predictions. After filtering using a p-value threshold of 0.9, 118 miRNAs and 118 false positives were returned thus giving a specificity of 50%. The 118 false positive

predictions that did not come from the same genomic loci were then searched against all *Arabidopsis* miRNA precursor sequences in miRBase. Out of these, 60 mapped to known miRNA hairpins. This can be explained by the fact that miRNAs often belong to families and have several identical or near identical copies spread across the genome. As this method has no concept of the genomic positions of sRNAs it is prone to predicting miRNA/miRNA* pairs which, even though they belong to the same family, come from different genomic loci. A summary of the full results are shown Table D.1.

6.5 Testing

In order to test the method using different organisms, both tomato (from Chapter 5) and rice (taken from CSRDB [Johnson et al., 2007]) 454 datasets were used. Perl code implementing the above method has been made available at `http://www.uea.ac.uk/~simonm/nogenome` and requires FASTA [Pearson, 2000], `RNAfold` [Hofacker, 2003] and `LIBSVM` [Chang and Lin, 2001] in order to function.

6.5.1 Rice analysis

Due to the relatively small size of the rice 454 dataset (11,809 non-redundant reads), all sequences of between 20-22nt with an abundance of 8 or more (186 non-redundant sequences) were used as input. The 186 sequences were then filtered to remove known non-coding RNAs leaving 121 unique sRNA sequences. Each of the input sequences were then used to search the total non-redundant sRNA sequence set in order to identify potential miRNA* sequences. Features of each potential pair were

then calculated as described in the Methods section.

In total 10,083 pairs were identified and of these the SVM classified 63 as being miRNA/miRNA* duplexes. After applying a p-value cutoff of 0.9, 37 predictions remained. Of the 37 putative miRNA/miRNA* pairs classified, 16 were derived from known miRNA loci and 21 false positives were found. Of the 21 false-positive predictions 12 were known miRNAs but the predicted star sequence was derived from a different locus (the miRNA* was from a different member of the same miRNA family). A summary of the full results are shown Table D.2.

6.5.2 Tomato analysis

As discussed in Chapter 5, the lack of a complete genome sequence has been a limiting factor in the analysis of tomato sRNAs. We therefore employed this approach in order to try to identify further novel miRNAs from the high-throughput sRNA dataset described in Chapter 5. sRNAs of 20-22nt with an abundance of at least 20 were extracted from the total sRNA dataset, yielding 811 non-redundant sRNAs (715 after filtering against non-coding RNA databases). These were then searched against the entire non-redundant sRNA dataset of 312,899 sequences and features were extracted as above yielding 265,867 potential pairs. Of these 422 were classified as miRNAs. Again a p-value threshold was applied and those pairs where p was below 0.9 were filtered out leaving 220 predicted miRNAs. A summary of the full results are shown Table D.3.

As a limited amount of genome sequence is available it is not possible to calculate the false positive rate for this dataset. However, 69 predicted miRNAs could be mapped to known plant hairpin precursors from miRBase and several interesting candidates were found. One such candidate pair could be mapped to an EST sequence present in the EMBL Nucleotide Sequence Database [Cochrane et al., 2006]. The sequence could not be folded into a miRNA hairpin using traditional folding algorithms and the miRNA (abundance 1349) and miRNA* (abundance 14) were 654 nt apart. The presence of such high abundance sRNAs in such close proximity appeared extremely unlikely, so this candidate was experimentally tested.

The presence of the predicted miRNA was first confirmed by Northern blot (Figure. 6.1) and showed differential expression in different developmental and tissue samples in tomato. Next it was confirmed that the sequence contained an intron (Figure. 6.2) which when spliced out caused the sequence to form a classical miRNA hairpin sequence (Figure. 6.3). Several other candidates were found during this analysis and will be tested by members of the Dalmay Laboratory.

6.6 Discussion

Methods to process sRNA datasets without a genome sequence are now essential as high-throughput sRNA sequencing is being used on a wide-spectrum of non-model organisms. It is of course possible to identify homologues of known/conserved miRNAs using sequence similarity searches such as BLAST [Altschul et al., 1990] and

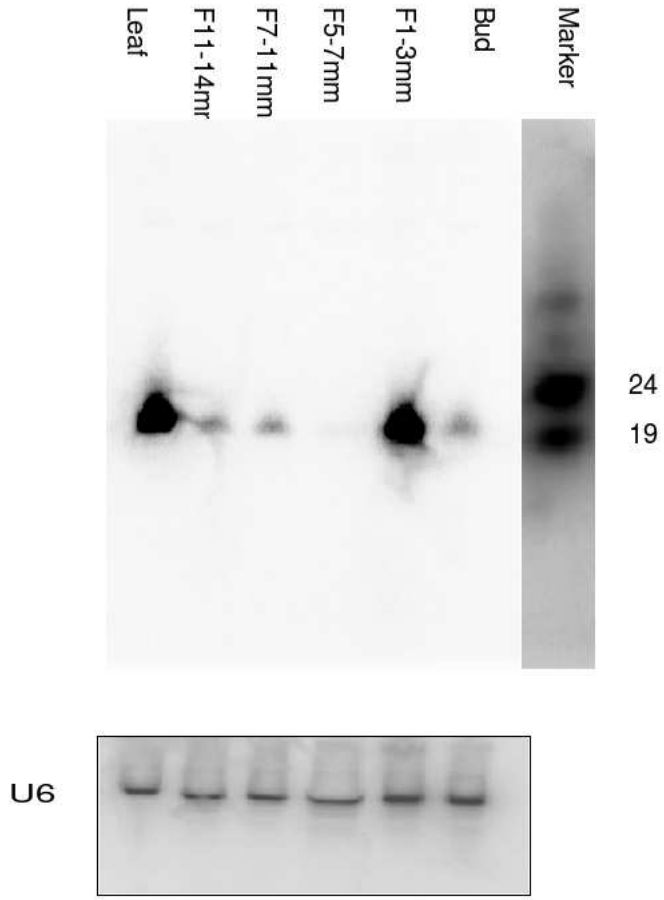

Figure 6.1: Northern blot of candidate miRNA from different plant tissues (leaf, fruit 11-14mm, fruit 7-11mm, fruit 5-7mm, fruit 1-3mm, bud).

FASTA [Pearson, 2000], but no tools exist to identify novel miRNAs. This approach, is more prone to false positives than traditional genome-centric methods but does have potential benefits even in cases where a genome sequence is available, since as well as being able to identify miRNAs detectable using traditional structure prediction and analysis tools (e.g. `miRCat` and `miRDeep`), it can potentially find spliced miRNAs (Figure. 6.3), nat-miRNAs [Lu et al., 2008] and other miRNAs which would not be picked up due to poor secondary structure or extremely long loop regions (e.g.

Figure 6.2: Alignment of spliced and un-spliced miRNA precursor “TOP14”. Splice variant TOP14_sv gives rise to a valid hairpin (see Figure. 6.3) structure whereas the un-spliced transcript TOP14 (see Figure. 6.4) does not. Identical regions are highlighted and the intron is un-coloured.

miR824 Figure. 6.5). The drawbacks of this method, in addition to the relatively high false-positive rate, is the fact that if a miRNA* is not present, then the miRNA cannot be predicted. This is especially problematic when the abundance of a miRNA is low since there is little chance of cloning a miRNA* sequence.

This method is currently the only technique that can be used to find miRNAs without a sequenced genome, and so it provides a useful tool for biologists who are attempting to find miRNAs in organisms with unsequenced genomes. The false positive rate of around 50%, is relatively high but is still acceptable for biologists to validate predictions experimentally. In addition the list can be sorted by abundance or p-value

to obtain the best candidates. A separate laboratory-based technique to sequence the full hairpin precursors from miRNA/miRNA* pairs is currently under development by members of the Dalmay Laboratory. A combined bioinformatics/experimental approach should lead to a powerful technique for finding new miRNAs in unsequenced genomes.

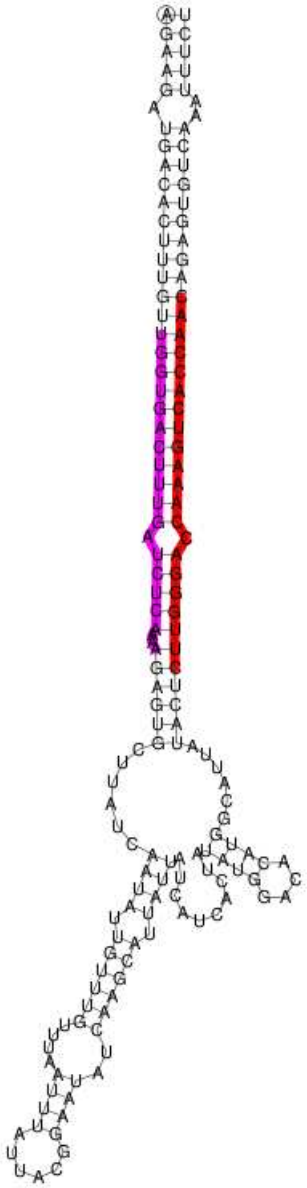

Figure 6.3: Predicted hairpin structure of the TOP14 pre-miRNA, miRNA is highlighted in red and the miRNA* sequence in pink.

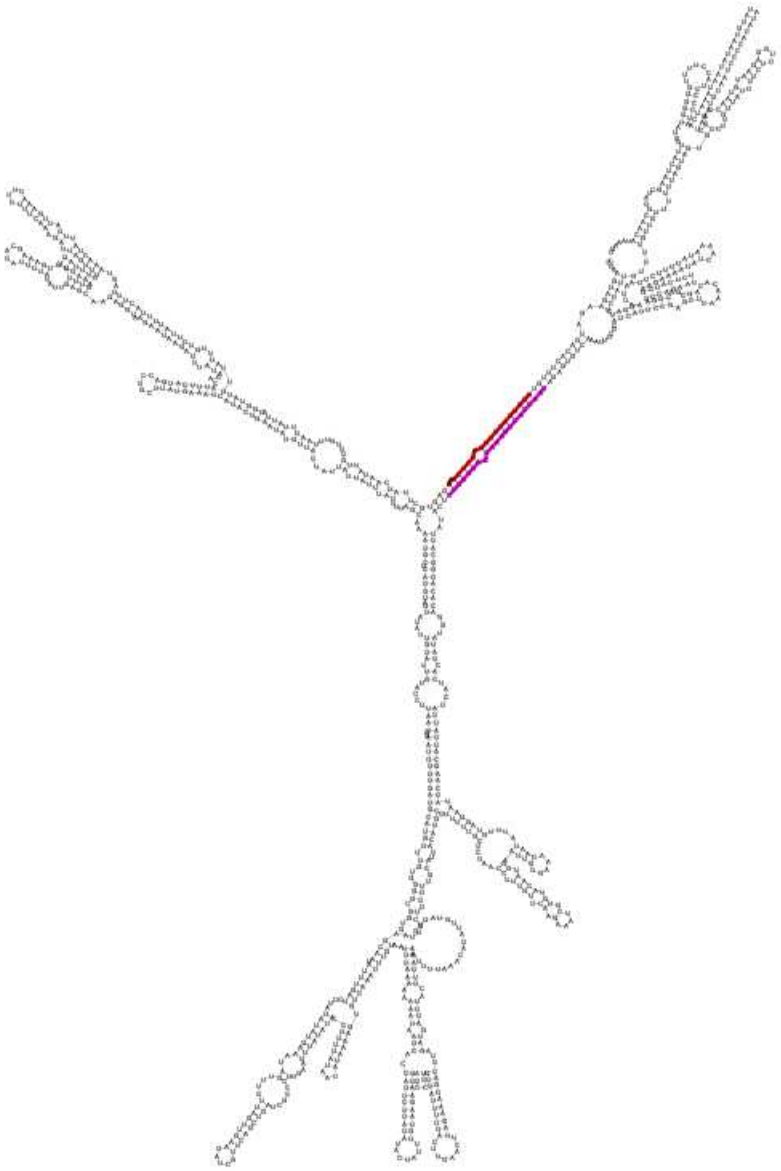

Figure 6.4: Predicted secondary structure of the unspliced transcript, “TOP14” miRNA is highlighted in pink with the miRNA* sequence in red. No valid miRNA hairpin structure is formed when the intron is not spliced out.

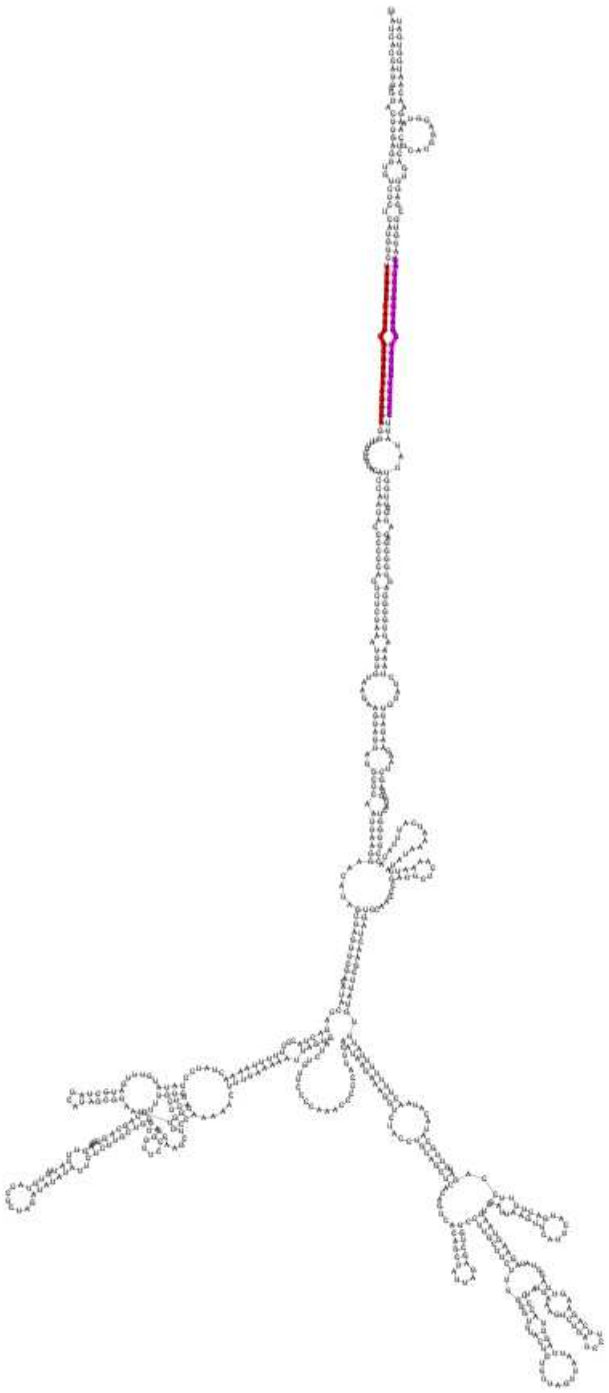

Figure 6.5: Predicted secondary structure of miR824 showing a non-typical secondary structure which would not be classified as a valid miRNA precursor by structure based methods such as `miRCat` yet is found using the SVM-based no genome prediction. miR824 is highlighted in red and miR824* is highlighted in pink.

Chapter 7

A scoring matrix approach to detecting miRNA target sites

This chapter is an adapted and extended version of

Moxon S., Moulton V., Kim J.T. (2008): A scoring matrix approach to detecting miRNA target sites. *Algorithms Mol Biol*. 3(1):3.

7.1 Summary

This chapter outlines a new approach to detecting miRNA targets that is applicable to both plants and animals. The method allows the user to take advantage of existing biological knowledge by incorporating previously validated targets for a particular miRNA into a search for novel target sites. We test the method using known miRNA targets and show that it performs well in terms of sensitivity and specificity in comparison to other methods.

7.2 Background

As mentioned in Chapter 2, animal miRNA target detection is a very difficult problem due to the poor complementarity between miRNA and target. Several computational methods have been developed for miRNA target prediction – see e.g. [Enright et al., 2003, Zhang, 2005, Krüger and Rehmsmeier, 2006, Mazière and Enright, 2007]. These methods usually rely on finding target sequences based on a single miRNA input, and employ nucleotide complementarity and MFE calculations to identify candidate miRNA/target duplexes. Although these methods have been successfully used in target prediction e.g. [John et al., 2004], their specificity can be limited, i.e. they may produce many false positives [Rajewsky, 2006].

Various methods have been proposed to improve the specificity of miRNA target prediction methods. For example, comparative genomics has been used to focus on sites that are conserved between species [John et al., 2004]. Here we concentrate

on an alternative approach, the Stacking Binding Matrix (*SBM*), in which we can incorporate all of the known targets for a given miRNA (in general a miRNA may target several sites) into a search for additional targets. The number of experimentally validated miRNA targets is steadily growing, and as this number increases so too should the usefulness of the *SBM* method.

7.3 Methods

Our approach is an adaptation of the binding matrix (BM) technique for transcription factor binding site classification [Kim et al., 2004], a method that was designed to systematically maximise specificity in searches for transcription factor binding sites. In contrast to computation of the BM, which uses single nucleotide information and results in a $4 \times l$ matrix for scoring words of length l, the *SBM* is a $16 \times (l-1)$ matrix based on dinucleotides (i.e. consecutive pairs of nucleotides). In this way, it is possible to incorporate the fundamental principle of RNA stacking energies [Turner et al., 1987] which is commonly used in miRNA detection.

In brief, the *SBM* is computed from a multiple sequence alignment consisting of the reverse complement of the miRNA in question together with any known target sequences. The resulting matrix (or set of matrices in case the alignment contains gaps) is then used to scan and score a set of potential target sequences. Sequences having a score exceeding a user-defined threshold are returned as potential targets.

7.3.1 Scoring matrices and the binding matrix

A *scoring matrix* for nucleotide words of length l is an $\{A, C, G, U\} \times l$ matrix $M = (m_{bk})$. Given a word $w = w[1]w[2] \ldots w[l]$ in the alphabet $\{A, C, G, U\}$ its score $S(w)$ is the sum of the matrix elements "selected" by the symbols in the word, that is,

$$S(w) = \sum_{k=1}^{l} m_{w[k],k}.$$

Given a threshold $S_{\min}$, a word w is classified as a *binding word* if $S(w) \geq S_{\min}$ and otherwise it is classified as a non-binding word. Generally, the threshold can be used to adjust sensitivity and specificity of classification: Assuming a positive correlation between density of true positives and score, lowering the threshold increases sensitivity and decreases specificity. Also, notice that for any $\lambda > 0$, scoring a word with the matrix λM and using the threshold $\lambda S_{\min}$ results in the same classification. A matrix classifier is called *consistent* with a set $B = \{b_1, \ldots, b_N\}$, of known binding words if it classifies them all correctly [Wolff et al., 2003], i.e. if $S(b) \geq S_{\min}$ for all $b \in B$.

There are various ways of constructing a scoring matrix from a set of binding words [Stormo, 2000]. The *Binding Matrix (BM)* is defined to be the matrix for which the number of words classified as binding words is minimal, under the condition that it is consistent. A method for computing the BM and a discussion of its properties is given in [Kim et al., 2004].

7.3.2 Incorporating stacking into binding matrix computations

A key feature in RNA structure prediction is the incorporation of stacking energies [Turner et al., 1987]. So as to capture information from both nucleotide complementarity and base pair stacking energies, in the computation of the *SBM* we score *dinucleotides*. Formally, for nucleotide words of length l, *SBM* is a $\{A, C, G, U\}^2 \times (l-1)$ matrix. It is computed by first converting each word w into the sequence $w[1]w[2], w[2]w[3], \ldots, w[l-1]w[l]$ of dinucleotides in $\{A, C, G, U\}^2$ and then optimising as with the BM. An example *SBM* as calculated from the alignment shown in Figure 7.1 is given in Table 7.1.

For performance reasons, to compute the *SBM* we use the optimisation approach described in [Madany Mamlouk et al., 2003] rather than the quadratic programming technique used in [Kim et al., 2004]. All *SBM*s are scaled so that a threshold of 1 corresponds to the most specific consistent classifier.

Note that in contrast to transcription factors, where only binding site sequences (binding words) are available, the reverse complement of the miRNA sequence itself provides information about the accepted target site sequences. Thus we include the reverse complement of the miRNA within the alignment of the known target sites.

7.3.3 Incorporating gaps

The complementarity of a miRNA binding to a target site is usually imperfect and commonly involves bulges (see Figure 7.2), which results in gapped alignments.

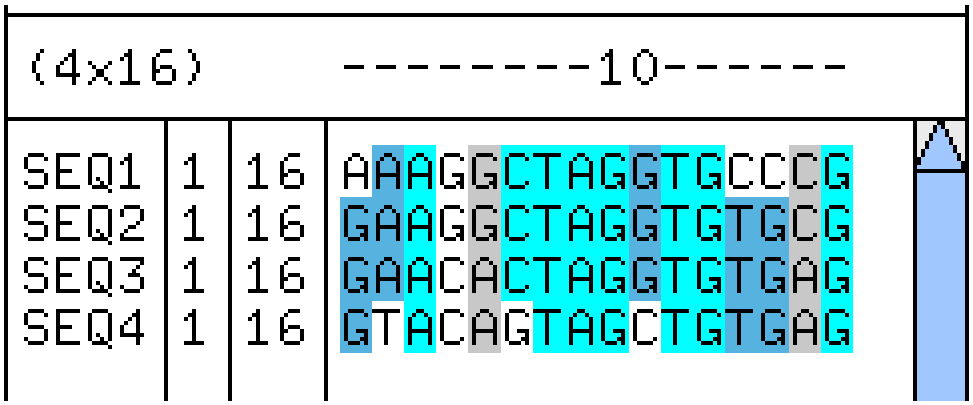

Figure 7.1: Screenshot of example input aligmment used to build the *SBM* in Table 7.1 as viewed using the Belvu alignment viewer (`http://sonnhammer.sbc.su.se/Belvu.html`). Columns are coloured based on nucleotide conservation using the default Belvu colour scheme.

Dinucleotide	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
aa:	0.04	0.07	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ac:	0.0	0.0	0.06	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ag:	0.0	0.0	0.06	0.0	0.05	0.0	0.0	0.11	0.0	0.0	0.0	0.0	0.0	0.0	0.06
at:	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ca:	0.0	0.0	0.0	0.06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
cc:	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.04	0.04	0.0
cg:	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.06
ct:	0.0	0.0	0.0	0.0	0.0	0.07	0.0	0.0	0.0	0.05	0.0	0.0	0.0	0.0	0.0
ga:	0.02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.06	0.0
gc:	0.0	0.0	0.0	0.0	0.06	0.0	0.0	0.0	0.05	0.0	0.0	0.04	0.0	0.01	0.0
gg:	0.0	0.0	0.0	0.06	0.0	0.0	0.0	0.0	0.07	0.0	0.0	0.0	0.0	0.0	0.0
gt:	0.05	0.0	0.0	0.0	0.0	0.05	0.0	0.0	0.0	0.07	0.0	0.07	0.0	0.0	0.0
ta:	0.0	0.04	0.0	0.0	0.0	0.0	0.11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
tc:	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
tg:	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.11	0.0	0.07	0.0	0.0
tt:	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 7.1: Example of a *SBM* scoring matrix from alignment given in Figure 7.1: The first column “Dinucleotide” shows each of the possible dinucleotide alignments. Each subsequence column shows the dinucleotide weighting (given to a maximum of two decimal places) as calculated from the input alignment (Figure 7.1).

However, in common with scoring matrix-based classification approaches, the *SBM* cannot accommodate gaps directly. To address this, we employ a *set* of *SBM*s rather than a single *SBM*.

For $N = \{A, C, G, U\}$, let $A = \{S_1, S_2, \ldots, S_n\}$ denote an alignment consisting of

```
ab 3′ UTR  5′ -          AGAA    A       - 3′
               -UUGU-CAG    UUGA UACCUCA
                :|:| |||:    :|:| |||||||
                GAUA GUU     GAUG AUGGAGU
let-7      3′ -U    U   G---    -       - 5′
```

Figure 7.2: Alignment of the *Drosophila melanogaster* let-7 miRNA to a cognate target site in the 3' UTR of the ab gene adapted from [Burgler and Macdonald, 2005, Fig. 1].

(possibly) gapped sequences over N of length l. Denote the gap character by $-$, and let $s_{i,j}$ be the j-th symbol of S_i. Suppose that $D \subseteq \{1, 2, \ldots, l\}$. Given a sequence $S_i \in A$, let $S_i^D = s_{i,j_1} s_{i,j_2} \ldots s_{i,j_{l-|D|}}$ denote the subsequence of S_i with $j_k < j_{k+1}$ and $j_k \in \{1, 2, \ldots, l\} - D$, and define the *subsequence alignment* of A corresponding to D to be $A^D = \{S_1^D, S_2^D, \ldots, S_n^D\}$ (i.e. the alignment obtained from A by removing the columns indexed by elements of D).

The *gap pattern of a sequence* $S_i \in A$, denoted $G(S_i)$, is the set $G(S_i) = \{j : s_{i,j} = -\}$. In particular, for each $S_i \in A$, the ungapped sequence corresponding to S_i equals $S_i^{G(S_i)}$. Correspondingly, the *gap pattern of* A is defined as $G(A) = \bigcup_i G(S_i)$, i.e. the set of indices of those columns in A that contain at least one gap.

Now, let $\mathcal{D}$ be a subset of $2^{G(A)}$ (in practice we take either $\mathcal{D}_{\text{all}} = 2^{G(A)}$ or $\mathcal{D}_{\text{observed}} = \{G(S) : S \in A\}$). For each of the alignments $\mathcal{A}(\mathcal{D}) = \{A^D : D \in \mathcal{D}\}$ we calculate a *SBM*. In case an alignment A' in $\mathcal{A}$ contains some gaps, each sequence S in A' that contains gaps is replaced by the set of all sequences obtained by replacing the gaps in S with all possible nucleotide symbol combinations (or the set of

nucleotides actually observed at the gap containing position).

Once the set of *SBM*s has been computed for each alignment in $\mathcal{A}(\mathcal{D})$, query sequences are then scanned with each of the matrices, and the final score at a given base in a query sequence is taken to be the largest of the scores attained by the individual *SBM*s. As usual, a target site is predicted in case the final score exceeds a user-defined threshold.

This extension to gapped alignments allows the detection of target sites with varying lengths whilst preserving specificity and consistency, both of which are key features of the original BM approach. Note that consistency is ensured since, for each sequence $S_i \in A$, we have $G(S_i) \in \mathcal{D}$ as one alignment in $\mathcal{A}(\mathcal{D})$ must contain $S_i^{G(S_i)}$. Computing *SBM*s based on $\mathcal{D}_{\text{observed}}$ makes most use of the gap information contained in the alignment. As an alternative, computing a (larger) *SBM* set based on $\mathcal{D}_{\text{all}}$ may allow detection of target sites that are recognised by a pairing structure different from those formed by the target sites known so far, which may be used to improve sensitivity.

7.3.4 Computational complexity

The number of alignments in the set $\mathcal{A}(\mathcal{D})$ used in the calculation of *SBM* set is of order $2^{|G(A)|}$, and so grows exponentially with the number of columns in A containing gaps. Hence, our approach will not scale to long alignments containing many gaps. Even so, in practice we have found the approach to be applicable to miRNA target prediction, since usually $|G(A)| \leq 6$ (as miRNAs are about 21nt's in length), resulting in at most $2^6 = 64$ alignments in $\mathcal{A}(\mathcal{D})$. Obviously, choosing $\mathcal{D} = \mathcal{D}_{\text{observed}}$ rather than

$\mathcal{D} = \mathcal{D}_{\text{all}}$ can considerably reduce $|\mathcal{D}|$, particularly if gaps occur in only a few distinct patterns. Likewise, the number of alignments obtained after the gap filling procedure is performed also grows exponentially, although the approach is still feasible for miRNA targets, again due to their short length.

7.3.5 Implementation

We have implemented our method in Python `http://www.python.org/` and R [R Development Core Team, 2004]. The code, together with documentation and examples, is freely available for download from `http://www.cmp.uea.ac.uk/~jtk/stackbm/`.

7.4 Results

To demonstrate the utility of the *SBM* method, we present an application to the problem of miRNA target detection for nematode worm (*Caenorhabditis elegans*), fruit fly (*Drosophila melanogaster*), mouse (*Mus musculus*), human (*Homo sapiens*) and thale cress (*Arabidopsis thaliana*). We also present a leave one out analysis, and a comparison with `miRanda` [Enright et al., 2003], a commonly used miRNA target prediction algorithm.

7.4.1 Data

We extracted *C. elegans*, *D. melanogaster*, *M. musculus* and *H. sapiens* miRNA entries from the miRBase database, release 9.1 [Griffiths-Jones et al., 2006] that

had more than one unique, experimentally validated target in the TarBase database [Sethupathy et al., 2006].

The reverse complement of each miRNA was then aligned with its validated target regions using the `ClustalW` alignment package [Chenna et al., 2003]. If local alignment algorithms are used, terminal gaps carry much less significance than internal gaps. Therefore, alignments were trimmed by removing columns containing terminal gaps at the 5' or 3' end. *SBM*sets were computed for these alignments as described in the Methods section. The *SBM* sets were used to search for potential new targets in the UTR sequence sets obtained from BioMart [Kasprzyk et al., 2004] (see Table 7.2 for details).

Organism	No. Sequences	Sequence type	No. Nucleotides
C. elegans	12,172	UTR	2,724,326
D. melanogaster	11,277	UTR	4,612,168
M. musculus	20,271	UTR	20,009,781
H. sapiens	27,685	UTR	30,673,888
A. thaliana	31,527	cDNA	46,447,255

Table 7.2: Summary of UTR datasets: "No. sequences" gives total number of unique sequences in this dataset; "Sequence type" gives the sequence type used (UTR or cDNA); "No. nucleotides" gives total number of nucleotides in the UTR set.

To test the applicability of the method to plant target prediction, we took a selection of *A. thaliana* miRNAs from miRBase together with validated target regions from the the *Arabidopsis* Small RNA Project Database (ASRP) [Gustafson et al., 2005], aligned these sequences with `ClustalW`, and computed *SBM*s.

7.4.2 Summary of *SBM* Scan

On the animal data sets, we determined for each of the *SBM* sets the number of predicted targets obtained by scanning the UTR data set, using a score threshold of 1. As in [Kim et al., 2004], we used the number of predicted targets obtained with a consistent classifier as an indicator of the classifier's specificity.

Plant miRNA targets usually occur in the protein coding region of genes and therefore we searched the gene sequence set TAIR6_cdna_20051108 obtained from The *Arabidopsis* Information Resource (TAIR) [Rhee et al., 2003] again using a threshold of 1. A summary of these results can be seen in Table 7.3.

In accordance with the definition of the *SBM* method, in Table 7.3 we see that all validated targets present in the input alignment are recovered in the scan output using a threshold of 1. In many cases no additional candidate targets are predicted using this stringent threshold, especially when there are few sequences provided in the input *SBM* set. Larger sets of validated targets tended to result in the prediction of more new candidate target sites, as illustrated in Table 7.3 by the cases of *dme-miR-4*, *dme-miR-7*, *cel-let-7* and *cel-miR-84*. This reflects the consistency criterion built into the binding matrix definition; a larger input set of sequences generally tended to reduce the stringency of the classifier.

cel-let-7 returned 1708 predicted targets at threshold 1 which appears to be relatively high compared with the other results, but given the size the searched database (2,274,326nt) it is a small proportion of all possible target regions. A possible reason

Organism	miRNA	Validated targets	Recovered targets	Potential novel targets
C. elegans	*cel-miR-273*	2	2	0
C. elegans	*cel-let-7*	15	15	1708
C. elegans	*cel-miR-84*	7	7	123
D. melanogaster	*dme-miR-11*	4	4	0
D. melanogaster	*dme-miR-2*	4	4	0
D. melanogaster	*dme-miR-4*	8	8	23
D. melanogaster	*dme-miR-7*	15	15	28
M. musculus	*mmu-miR-124*	3	3	0
M. musculus	*mmu-miR-206*	3	3	0
H. sapiens	*hsa-miR-1*	4	4	0
H. sapiens	*hsa-miR-122*	3	3	0
A. thaliana	*ath-miR-163*	5	5	0
A. thaliana	*ath-miR-172*	6	6	0
A. thaliana	*ath-miR-390*	1	1	0
A. thaliana	*ath-miR-398*	2	2	0
A. thaliana	*ath-miR-408*	2	3	1

Table 7.3: *SBM* scan summary obtained using a score threshold of 1: “miRNA” gives miRBase miRNA identifier; “Validated targets” gives number of unique validated targets present in the starting alignment; “Recovered targets” gives number of validated targets in the input alignment that were recovered; “Predicted novel targets” gives number of candidate target sequences (other than the validated targets) predicted by the *SBM* method.

for the large number of predicted targets is that the input sequence set used to build the *SBM* set was misaligned by `ClustalW`. The validated targets used to create the alignment showed a greater degree of heterogeneity that those in other alignments. Another possible explanation is that *cel-let-7* is known to have several paralogs (*cel-miR-84*, *cel-miR-48* and *cel-miR-241*) [Hayes et al., 2006] and therefore its targets are likely to overlap with other members of this miRNA family. It has also been suggested that some miRNAs may target thousands of different genes [Miranda et al., 2006] making it possible that many of the targets predicted are in fact true positives.

7.4.3 Leave one out analysis

While the *SBM* method used with $S_{\min} = 1$ recovers all targets that are present in the input alignment, unknown targets that receive a score below 1 are likely to exist. It is possible to detect such sequences using the *SBM* method by lowering the threshold. This increases the classifier's sensitivity at the expense of reducing its specificity. To assess this effect quantitatively we conducted a leave one out analysis. In particular we constructed leave one out alignments by deleting one target site sequence from an input alignment. Then, for each alignment in which the target sequence w was left out, we computed a *SBM* set and determined the score $S(w)$ of the target site that was left out. If $S(w) < 1$, the threshold needs to be adjusted to $S_{\min} = S(w)$ in order to detect w. We therefore scanned the respective UTR set with $S_{\min} = \min\{1, S(w)\}$ and determined the number of predicted targets.

An input alignment of n sequences allows construction of $n - 1$ leave one out alignments (we did not leave out the reverse complement of the miRNA), so data sets containing more experimentally validated target sites clearly result in more meaningful leave one out analyses. We therefore chose the four miRNAs that had the greatest number of known experimentally validated targets; *D. melanogaster miR-7* and *C. elegans let-7*, which both targeted 15 unique UTR regions as well as *D. melanogaster miR-4* (8 unique targets) and *C. elegans miR-84* (7 unique targets).

In total 2,484,850 UTR regions were scanned in the *C. elegans* set compared with to 4,409,641 regions in the *D. melanogaster* set. The score of each left out target

along with the number of regions with a score equal to or greater than this value in the scan using the full alignment are shown in Tables 7.4 and 7.5.

***Drosophila melanogaster*, miR-7**		
Target	**LOO score**	**$\geqslant$ LOO score**
CG12487.3/223-241	0.946	94
CG5185.3/279-297	1.000	34
CG3096.3/152-170	1.000	34
CG12487.3/250-268	1.000	34
CG3166.3/1100-1118	0.951	76
CG6096.3/103-121	1.000	34
CG8346.3/78-96	0.966	58
CG5185.3/334-352	1.000	34
CG6494.3/447-465	0.919	155
CG6096.3/24-42	1.000	34
CG6096.3/68-86	0.961	65
CG8328.3/63-81	0.773	2015
CG3166.3/1586-1602	0.855	393
CG3166.3/29-46	0.845	513
CG3166.3/1294-1312	0.861	521
***Caenorhabditis elegans*, let-7**		
Target	**LOO score**	**$\geqslant$ LOO score**
ZK792.6/247-264	0.959	3561
F38A6.1a/271-288	1.000	1708
C18D1.1.1/526-542	0.906	10458
ZK792.6/666-683	0.959	3522
ZK792.6/458-475	0.929	7311
F38A6.1a/133-150	0.874	19177
C01G8.9a/21-38	0.850	23906
ZK792.6/132-148	0.859	20570
C01G8.9a/159-175	0.813	30895
ZK792.6/190-207	0.807	41812
C12C8.3a/693-709	0.791	39369
C12C8.3a/742-757	1.000	1499
ZK792.6/484-499	0.898	10232
F11A1.3a/1007-1021	0.948	4658
ZK792.6/343-361	0.955	4352

Table 7.4: Leave one out analysis for *dme-miR-7 & cel-let-7*: "target" gives validated target sequence accession/start-end; "miRNA" gives miRNA targeting that region; "$\geqslant$ LOO score" gives mean number of regions scoring equal to or greater than the left out sequence.

The *SBM* method appears to show a greater degree of accuracy in the *D. melanogaster* miR-7 results. Here the mean score of the left out target is 0.9385

***Drosophila melanogaster*, miR-4**		
Target	**LOO score**	**$\geqslant$ LOO score**
CG6096.3/135-154	0.755	3118
CG8328.3/27-45	1.000	8
CG3096.3/33-52	0.929	161
CG3096.3/138-157	0.877	473
CG5185.3/46-65	0.960	64
CG12487.3/188-208	0.820	1298
CG12487.3/62-82	0.871	627
CG6096.3/210-230	0.908	207
***Caenorhabditis elegans*, miR-84**		
Target	**LOO score**	**$\geqslant$ LOO score**
ZK792.6/126-148	0.804	4970
ZK792.6/187-207	0.552	132626
ZK792.6/249-264	0.947	355
ZK792.6/342-361	0.761	12552
ZK792.6/460-475	0.858	2012
ZK792.6/479-499	0.739	18375
ZK792.6/665-683	0.726	15846

Table 7.5: Leave one out analysis for *dme-miR-4* & *cel-miR-84*: "target" gives validated target sequence accession/start-end; "miRNA" gives miRNA targeting that region; "$\geqslant$ LOO score" gives mean number of regions scoring equal to or greater than the left out sequence.

and the mean number of target regions scoring greater than or equal to the left out sequence is 273 (0.006% of the total UTR regions scanned). The *C. elegans* let-7 scan indicates a lower degree of specificity, with an average score of 0.9032, returning a mean of 14869 regions with a score greater than or equal to the score of the left out validated target sequence. This represents 0.598% of the sequence database that was searched.

For *D. melanogaster* miR-4 the *SBM* method gave a mean score of 0.890 with an average of 745 target regions scoring greater than or equal to the left out sequence (0.017% of the total UTR regions scanned), and for *C. elegans* miR-84 a mean score of 0.770 was obtained and an average of 26677 target regions scoring greater than

or equal to the left out sequence was returned (1.074% of the total UTR regions scanned). The decrease in specificity in the *C. elegans* miR-84 results is largely due to a single leave one out test in which over 132626 sequences scored higher than the left out sequence (which received a score of 0.552).

Overall, the lowering of the threshold required to detect a word not in the input set results in a moderate increase in the number of reported hits, which is indicative of a high specificity even with the reduced threshold.

In order to assess the performance of the algorithm when few known targets are provided in the input alignment we re-ran the *C. elegans let-7* and *D. melanogaster miR-7* scans but this time split each of the alignments of 15 validated targets into two subalignments containing 8 and 7 sequences respectively. Table 7.6 shows that as the number of sequences used to build the *SBM* decreases, so does the mean score of the left out sequences. This indicates, as might be expected that as the number of sequences left out of the alignment increases the specificity decreases.

	15 targets	**14 targets**	**8 targets**	**7 targets**
Mean score *C. elegans let-7*	1.000	0.903	0.851	0.810
Mean number returned *C. elegans let-7*	1708	14869	18032	17225
Mean score *D. melanogaster miR-7*	1.000	0.938	0.908	0.890
Mean number returned *D. melanogaster miR-7*	28	273	509	138

Table 7.6: Leave several out analysis: Shows mean scores and mean number of regions scoring above maximal consistent threshold for alignments containing 15, 14, 8 and 7 validated targets.

7.4.4 Comparison with miRanda

We also compared the performance of the *SBM* method with `miRanda` v1.9, a commonly used target prediction tool [Enright et al., 2003]. `miRanda` takes a single miRNA sequence as input and searches a sequence dataset for potential target regions. It uses two different criteria to detect potential target sites, the alignment score and the MFE of the miRNA bound to the potential target sequence.

In order to obtain results with `miRanda` that could be meaningfully compared with the *SBM* method, we used `miRanda` to score every potential target site across each of the UTR sequences. To do this we split each of the UTRs into 30nt sequence windows covering the entire length of each UTR and used this as our sequence database for the `miRanda` scan. Since the same target region may be scored more than once using this approach, we removed any duplicate regions from the results before the comparison. By default `miRanda` uses relatively stringent threshold values which do not necessarily recover all known target regions, i.e. classification is not consistent. For this reason `miRanda` was run using a negative score threshold and a positive energy threshold which allowed us to obtain a wide distribution of scores and to ensure consistency.

Table 7.7 provides an overview of the `miRanda` comparison (full results can be found in Appendix E). In general the *SBM* method compared favourably with `miRanda`. This is not unexpected as we incorporate additional information into our searches. For example the *cel-let7* results show that an average of 14869 regions

had a score that was at least as high as the left out sequence using *SBM* whereas an average of 92332 regions scored at least as high as the validated target using `miRanda`. This difference was more pronounced in the *dme-miR-7* results where an average of 273 sequences scored equal to or better than the left out sequences and an average of 8868 sequences scored at least as high as the validated target using `miRanda`. The *SBM* method returned an average of 745 sequences scoring equal to or better than the left out sequence for *dme-miR-4* in comparison to an average of 11488 sequences that scored at least as high as the validated target using `miRanda`. An average of 26677 target regions were returned using the *SBM* method for *cel-miR-84* compared with 190693 using `miRanda`.

miRNA	LOO score	$\geqslant$ LOO score	miRanda(s)	$\geqslant$ miRanda(s)	miRanda(e)	$\geqslant$ miRanda(e)	$\geqslant$ miRanda(se)
cel-let-7	0.903	14869	119	92332	-15.46	60266	23992
cel-miR-84	0.770	26677	106	190693	-10.19	150137	48538
dme-miR-7	0.938	273	159	8868	-21.69	7227	2129
dme-miR-4	0.890	745	131	11488	-8.51	184134	5325

Table 7.7: Summary of results for the leave one out analysis: “miRNA” gives miRBase accession of the miRNA sequence; “LOO score” gives mean score of the targets left out of the *SBM*; “$\geqslant$ LOO score” gives mean number of regions scoring equal to or greater than the left out sequence; “miRanda(s)” gives raw score of the miRanda hit of lowest scoring target region; “$\geqslant$ miRanda(s)” gives number of regions with returned using the maximal consistent score threshold; “miRanda(e)” gives minimum free energy (MFE) of the miRanda hit of the least stable target region; “$\geqslant$ miRanda(e)” gives number of regions with returned using the maximal consistent MFE threshold; “$\geqslant$ miRanda(se)” gives number of regions with returned using the maximal consistent combined score and MFE threshold.

We determined the maximal consistent threshold for `miRanda` results by filtering out all candidates with an alignment score lower than the lowest scoring validated target. The remaining candidates are then filtered further by removing any sequence

with an MFE of greater than the MFE of the highest (least stable) of the validated targets. The number of regions returned using the maximal consistent threshold in `miRanda` were 23992 for *cel-let7* in contrast to the 1708 returned using the *SBM* method with maximal consistent threshold. 48538 regions were recovered for *cel-miR-84* compared with 123 using *SBM*, 2129 for *dme-miR-4* in comparison to 23 with *SBM* and 5325 for *dme-miR-7*, with the *SBM* method returning 28.

7.5 Discussion

We have presented a new method, *SBM*, that allows the use of miRNA target site sequences in addition to the miRNA sequence itself to search for novel target sites. We have demonstrated its application to target prediction for a variety of miRNA examples from different organisms and have shown that it performs well in comparison to `miRanda`.

Many computational methods for target prediction tend to suffer from a lack of specificity [Rajewsky, 2006]. The *SBM* method allows the use of all known target sequences in the search, and is designed to provide maximum specificity whilst recovering all members present in the starting alignment. Thus, as the number of experimentally validated miRNA targets grows, the *SBM* method should provide an attractive addition to the available miRNA target site detection methods.

Many current target prediction techniques are based on algorithms with fixed parameters (such as base pairing rules or binding energies) that are used to assess

potential targets by matching them to the miRNA sequence. These algorithms are designed to reflect molecular target recognition mechanisms that are assumed to apply to miRNA target recognition in general. Tailoring these algorithms to reflect mechanisms that are specific to the miRNA is difficult or impossible. In contrast to this, the *SBM* method can capture aspects of specific binding mechanisms by extracting such specific information from the set of validated target site sequences. This also makes the method generic in that it can be applied to any organism without having to assume any prior knowledge of specific target recognition mechanisms.

Due to the small number of validated targets for each miRNA, the maximal consistent threshold used in the *SBM* method is rather stringent. We chose this threshold to facilitate comparison of the method to `miRanda`. For many applications lowering thresholds to increase sensitivity at the cost of losing some specificity may be advisable. The specificity advantage of the *SBM* method can be expected to be partly independent of the threshold, since moderate relaxation of the threshold for a classifier that attains a high level of specificity with a given threshold can be assumed to retain some of the specificity advantage.

As with all scoring matrix approaches, the *SBM* method is limited by the quality of the input data. Firstly if a false positive target sequence is provided as input the method will be adversely affected, therefore only experimentally validated targets should generally be used as input. Secondly the quality of the input alignment is extremely important and a poor quality alignment will lead to poor performance. miRNAs are

relatively short (~21nt) which means that in many cases they can be aligned quite accurately using multiple alignment algorithms such as `ClustalW` [Chenna et al., 2003] and `MUSCLE` [Edgar, 2004]. In some cases however, the conservation between sites targeted by the same miRNA is very low, meaning that an accurate sequence alignment is hard to produce using automated methods. In such cases it may be favourable to hand curate alignments in order to ensure quality and obtain optimal *SBM* results.

Thirdly, although the short length of miRNAs also allows for the integration of gapped alignments in the *SBM* method, the method will only search for the gap patterns contained in the input alignment. Thus, if targets contain insertion/deletion patterns which are not specified in this way, then they may receive a lower score or even be missed completely depending on the threshold used in the search.

Several miRNA target prediction systems have implemented post-processing steps in order to increase their specificity. The most commonly used filtering approach is to look for cross-species conservation of target sites. Here target sites that appear not to be conserved between multiple species are filtered out from the search results, removing false positives, and leading to increased specificity. This type of approach could be applied to results obtained with *SBM* to further increase the specificity of target predictions. However, we note that this might also lead to a reduction in sensitivity as it is now known that miRNAs themselves are not always conserved between related species (e.g. [Fahlgren et al., 2007]). Another possibility is to post-process based on target

site accessibility. It has recently been shown that taking into account target site accessibility in the 3' UTR can improve target prediction accuracy [Kertesz et al., 2007]. For instance if a predicted target site is part of a stable secondary structure (and is therefore already involved in base-pairing) it is less likely that the miRNA will be able to bind to the target causing the translational repression of the mRNA.

In conclusion, we have presented a promising new method for miRNA target prediction, *SBM*, that employs a generic scoring matrix approach and incorporates experimentally validated targets. Since the number of validated targets is constantly growing, *SBM* should provide a useful new addition to the current target prediction toolbox.

Chapter 8

Conclusions and future work

8.1 Summary

In the previous chapters we have described some new tools for the analysis of sRNA data. In Chapter 3 we introduced `miRCat`, a tool for finding miRNAs in high-throughput sequence datasets, and then went on to demonstrate its application to both small-scale (Chapter 4) and high-throughput (Chapter 5) sRNA sequence datasets in tomato, as well as validating the method in *Arabidopsis* and extending its functionality to animals. We have implemented `miRCat` along with several other tools including those for target prediction and ta-siRNA classification on a web-server which is being used by the scientific community for the analysis of high-throughput sequence data.

We have also developed a new method for the classification of miRNAs from high-throughput sequence data without the need for a genome sequence (see Chapter 6). This method relies on detecting miRNA/miRNA* pairs from a sRNA sequence set and uses a SVM in order to classify real miRNA/miRNA* pairs. We demonstrated its use by finding both known and novel miRNAs and showed that, unlike other methods that

use genomic coordinates and secondary structure information, the SVM-based classifier can detect non-standard miRNAs such as those containing introns or exceedingly long loop regions. As a result of this we were able to identify a novel tomato miRNA precursor containing an intron.

In Chapter 7 we described *SBM*, a scoring matrix approach to miRNA target prediction. *SBM* is able to utilise existing, validated miRNA target information for a given miRNA in order to improve the specificity of future predictions. *SBM* is generic and does not rely on fixed rules that are generally applied to miRNA target prediction, so that the method can be applied to both animals and plants.

8.2 Future Work

Further extensions to some of the methods and tools presented in this thesis are possible and are discussed below.

8.2.1 Improvements to miRNA prediction in unsequenced genomes

The results obtained in Chapter 6 are preliminary and it is likely that the model used to train the SVM can be improved to obtain more accurate results by, for example, adding more features and using further training examples. The method could also be extended to animal datasets by training a separate model on animal miRNA/miRNA* pairs from high-throughput sequencing.

As with the `miRCat` tool, a web-based implementation of this tool would be useful

as it removes the need to install and run software locally. A user could then upload a dataset of interest and receive a ranked list of candidate miRNA sequences for follow up experimentation.

8.2.2 Improvements to the *SBM* method

The *SBM* method introduced a new concept in miRNA target recognition by using information about the miRNA and known, validated targets of the miRNA in the search for new target sites. The method was shown to increase search specificity but there are various potential improvements which could be made in order to further improve the results. For example, it would be interesting to add filters for both MFE (as used by `miRanda`) [John et al., 2004] and target site accessibility, used by Kertesz *et al*. [Kertesz et al., 2007], to try to improve the specificity of this method further.

8.3 Conclusions

High-throughput sequencing technologies have enabled us to obtain complex small RNA profiles of biological samples and has led to the discovery of a wealth of previously unknown miRNAs and other sRNAs over recent years. As seen from work presented in this thesis, computational tools are invaluable for the interpretation of such data and have led to a number of important discoveries such as a miRNA involved in fruit ripening and a new class of miRNA precursor which contains an intron and is only processed in its spliced form.

As high-throughput technologies mature and evolve it is likely that the number of

sequences obtained from a single sample will increase by an order of magnitude, and even now millions of sequence reads can be obtained from a single experiment. Future increases in sequencing depth alone are unlikely to lead to the discovery of many new miRNAs as it is likely that the majority of ubiquitously expressed miRNAs in previously studied model organisms such as *Arabidopsis*, human and *Drosophila* have been described. However, only a small number of commercially and medically important plants and animals have a fully sequenced genome so it has been impossible to look for novel miRNAs in such organisms. An important objective now is to provide computational methods which will allow us to identify miRNAs in unsequenced genomes such as the approach discussed in Chapter 6.

Another important challenge lies in understanding the role of miRNAs in the cell. This is especially true for animals where the vast majority of characterised miRNAs have no known function. A key component in the understanding miRNA function lies in the use of bioinformatics and the development of new, more accurate target prediction algorithms such as *SBM* that can be used to confidently predict the genes regulated by a given miRNA.

As high-throughput sequencing is rapidly becoming cheaper, it is now already economically viable to sequence several different samples from the same organism (e.g. different tissue types, developmental stages or a time-series after treatment). This multi-sample sequencing is likely to lead to the discovery of new miRNAs which are only expressed under specific conditions, developmental stages or tissue types. It also

means that sRNA profiles can be compared (e.g. over developmental stages) leading to the discovery of differentially expressed sRNA loci and giving an insight into sRNA function in biological processes.

Small RNA biology and bioinformatics are rapidly evolving due to the advent and evolution of high-throughput sequencing. Computational methods for the classification and analysis of these datasets, such as those presented in this thesis, have both supported and made possible exciting new discoveries in the field. Our knowledge and understanding of sRNAs has increased dramatically from 2001 and the characterisation of the first miRNA, to the present day. We are now aware of the complex and subtle mechanisms of gene regulation performed by an ever increasing number of different sRNA classes. It is likely, in this relatively new field of biology, that there are many new classes of sRNAs with which have not yet been characterised and the mechanisms and networks of sRNA regulation may be more complex than we first thought. Whatever discoveries the future holds it is clear that bioinformatics will play a key role in deciphering the sRNA content of the cell.

Bibliography

[Adai et al., 2005] Adai, A., Johnson, C., Mlotshwa, S., Archer-Evans, S., Manocha, V., Vance, V., and Sundaresan, V. (2005). Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res*, 15(1):78–91.

[Adams-Phillips et al., 2004] Adams-Phillips, L., Barry, C., Kannan, P., Leclercq, J., Bouzayen, M., and Giovannoni, J. (2004). Evidence that CTR1-mediated ethylene signal transduction in tomato is encoded by a multigene family whose members display distinct regulatory features. *Plant Mol Biol*, 54(3):387–404.

[Addo-Quaye et al., 2008] Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. *Curr Biol*, 18(10):758–762.

[Adenot et al., 2006] Adenot, X., Elmayan, T., Lauressergues, D., Boutet, S., Bouché, N., Gasciolli, V., and Vaucheret, H. (2006). DRB4-dependent TAS3 trans-acting siRNAs control leaf morphology through AGO7. *Curr Biol*, 16(9):927–932.

[Adé and Belzile, 1999] Adé, J. and Belzile, F. J. (1999). Hairpin elements, the first family of foldback transposons (FTs) in Arabidopsis thaliana. *Plant J*, 19(5):591–597.

[Allen et al., 2005] Allen, E., Xie, Z., Gustafson, A. M., and Carrington, J. C. (2005). microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell*, 121(2):207–221.

[Allen et al., 2004] Allen, E., Xie, Z., Gustafson, A. M., Sung, G.-H., Spatafora, J. W., and Carrington, J. C. (2004). Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. *Nat Genet*, 36(12):1282–1290.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

[Ambros et al., 2003] 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.

[Artzi et al., 2008] Artzi, S., Kiezun, A., and Shomron, N. (2008). miRNAminer: a tool for homologous microRNA gene search. *BMC Bioinformatics*, 9:39.

[Aukerman and Sakai, 2003] Aukerman, M. J. and Sakai, H. (2003). Regulation of flowering time and floral organ identity by a MicroRNA and its APETALA2-like target genes. *Plant Cell*, 15(11):2730–2741.

[Axtell and Bartel, 2005] Axtell, M. J. and Bartel, D. P. (2005). Antiquity of microRNAs and their targets in land plants. *Plant Cell*, 17(6):1658–1673.

[Axtell et al., 2006] Axtell, M. J., Jan, C., Rajagopalan, R., and Bartel, D. P. (2006). A two-hit trigger for siRNA biogenesis in plants. *Cell*, 127(3):565–577.

[Axtell et al., 2007] Axtell, M. J., Snyder, J. A., and Bartel, D. P. (2007). Common functions for diverse small RNAs of land plants. *Plant Cell*, 19(6):1750–1769.

[Baker et al., 2005] Baker, C. C., Sieber, P., Wellmer, F., and Meyerowitz, E. M. (2005). The early extra petals1 mutant uncovers a role for microRNA miR164c in regulating petal number in Arabidopsis. *Curr Biol*, 15(4):303–315.

[Barakat et al., 2007a] Barakat, A., Wall, K., Leebens-Mack, J., Wang, Y. J., Carlson, J. E., and Depamphilis, C. W. (2007a). Large-scale identification of microRNAs from a basal eudicot (Eschscholzia californica) and conservation in flowering plants. *Plant J*, 51(6):991–1003.

[Barakat et al., 2007b] Barakat, A., Wall, P. K., Diloreto, S., Depamphilis, C. W., and Carlson, J. E. (2007b). Conservation and divergence of microRNAs in Populus. *BMC Genomics*, 8:481.

[Barrett et al., 2007] Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I. F., Soboleva, A., Tomashevsky, M., and Edgar, R. (2007). NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*, 35(Database issue):D760–D765.

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

[Bennett, 2004] Bennett, S. (2004). Solexa Ltd. *Pharmacogenomics*, 5(4):433–438.

[Bentwich et al., 2005] Bentwich, I., Avniel, A., Karov, Y., Aharonov, R., Gilad, S., Barad, O., Barzilai, A., Einat, P., Einav, U., Meiri, E., Sharon, E., Spector, Y., and

Bentwich, Z. (2005). Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet*, 37(7):766–770.

[Berezikov et al., 2005] Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R. H. A., and Cuppen, E. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, 120(1):21–24.

[Bonnet et al., 2004a] Bonnet, E., Wuyts, J., Rouzé, P., and de Peer, Y. V. (2004a). Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proc Natl Acad Sci U S A*, 101(31):11511–11516.

[Bonnet et al., 2004b] Bonnet, E., Wuyts, J., Rouzé, P., and de Peer, Y. V. (2004b). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20(17):2911–2917.

[Borsani et al., 2005] Borsani, O., Zhu, J., Verslues, P. E., Sunkar, R., and Zhu, J.-K. (2005). Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. *Cell*, 123(7):1279–1291.

[Brennecke et al., 2007] Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., and Hannon, G. J. (2007). Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, 128(6):1089–1103.

[Brenner et al., 2000] Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D. H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S. R., Moon, K., Burcham, T., Pallas, M.,

DuBridge, R. B., Kirchner, J., Fearon, K., Mao, J., and Corcoran, K. (2000). Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol*, 18(6):630–634.

[Brodersen et al., 2008] Brodersen, P., Sakvarelidze-Achard, L., Bruun-Rasmussen, M., Dunoyer, P., Yamamoto, Y. Y., Sieburth, L., and Voinnet, O. (2008). Widespread translational inhibition by plant miRNAs and siRNAs. *Science*, 320(5880):1185–1190.

[Burgler and Macdonald, 2005] Burgler, C. and Macdonald, P. M. (2005). Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method. *BMC Genomics*, 6(1):88.

[Cai et al., 2005] Cai, X., Lu, S., Zhang, Z., Gonzalez, C. M., Damania, B., and Cullen, B. R. (2005). Kaposi's sarcoma-associated herpesvirus expresses an array of viral microRNAs in latently infected cells. *Proc Natl Acad Sci U S A*, 102(15):5570–5575.

[Calabrese et al., 2007] Calabrese, J. M., Seila, A. C., Yeo, G. W., and Sharp, P. A. (2007). RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. *Proc Natl Acad Sci U S A*, 104(46):18097–18102.

[Carbone et al., 2005] Carbone, F., Pizzichini, D., Giuliano, G., Rosati, C., and Perrotta, G. (2005). Comparative profiling of tomato fruits and leaves evidences a complex modulation of global transcript profiles. *Plant Sci*, 169:165–175.

[Carrari and Fernie, 2006] Carrari, F. and Fernie, A. R. (2006). Metabolic regulation underlying tomato fruit development. *J Exp Bot*, 57(9):1883–1897.

[Chan and Ding, 2008] Chan, C. Y. and Ding, Y. (2008). Boltzmann ensemble features of RNA secondary structures: a comparative analysis of biological RNA sequences and random shuffles. *J Math Biol*, 56(1-2):93–105.

[Chan et al., 2004] Chan, S. W.-L., Zilberman, D., Xie, Z., Johansen, L. K., Carrington, J. C., and Jacobsen, S. E. (2004). RNA silencing genes control de novo DNA methylation. *Science*, 303(5662):1336.

[Chang and Lin, 2001] Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Chapman and Carrington, 2007] Chapman, E. J. and Carrington, J. C. (2007). Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet*, 8(11):884–896.

[Chen et al., 2007] Chen, H.-M., Li, Y.-H., and Wu, S.-H. (2007). Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis. *Proc Natl Acad Sci U S A*, 104(9):3318–3323.

[Chen, 2004] Chen, X. (2004). A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science*, 303(5666):2022–2025.

[Chenna et al., 2003] Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T. J., Higgins, D. G., and Thompson, J. D. (2003). Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res*, 31(13):3497–3500.

[Clancy et al., 2007] Clancy, J. L., Nousch, M., Humphreys, D. T., Westman, B. J., Beilharz, T. H., and Preiss, T. (2007). Methods to analyze microRNA-mediated control of mRNA translation. *Methods Enzymol*, 431:83–111.

[Clote et al., 2005] Clote, P., Ferré, F., Kranakis, E., and Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–591.

[Cochrane et al., 2006] Cochrane, G., Aldebert, P., Althorpe, N., Andersson, M., Baker, W., Baldwin, A., Bates, K., Bhattacharyya, S., Browne, P., van den Broek, A., Castro, M., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Kanz, C., Kulikova, T., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., McHale, M., McWilliam, H., Mukherjee, G., Nardone, F., Pastor, M. P. G., Sobhany, S., Stoehr, P., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W., and Apweiler, R. (2006). EMBL Nucleotide Sequence Database: developments in 2005. *Nucleic Acids Res*, 34(Database issue):D10–D15.

[Combier et al., 2006] Combier, J.-P., Frugier, F., de Billy, F., Boualem, A., El-Yahyaoui, F., Moreau, S., Vernié, T., Ott, T., Gamas, P., Crespi, M., and Niebel, A. (2006). MtHAP2-1 is a key transcriptional regulator of symbiotic nodule development regulated by microRNA169 in Medicago truncatula. *Genes Dev*, 20(22):3084–3088.

[Dalmay et al., 1993] Dalmay, T., Rubino, L., Burgyán, J., Kollár, A., and Russo, M. (1993). Functional analysis of cymbidium ringspot virus genome. *Virology*, 194(2):697–704.

[Dezulian et al., 2006] Dezulian, T., Remmert, M., Palatnik, J. F., Weigel, D., and Huson, D. H. (2006). Identification of plant microRNA homologs. *Bioinformatics*, 22(3):359–360.

[Eddy, 2004] Eddy, S. R. (2004). What is dynamic programming? *Nat Biotechnol*, 22(7):909–910.

[Edgar, 2004] Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113.

[Enright et al., 2003] Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C., and Marks, D. S. (2003). MicroRNA targets in Drosophila. *Genome Biol*, 5(1):R1.

[Fahlgren et al., 2007] Fahlgren, N., Howell, M. D., Kasschau, K. D., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., Givan, S. A., Law, T. F., Grant, S. R., Dangl, J. L., and Carrington, J. C. (2007). High-Throughput Sequencing of Arabidopsis microRNAs: Evidence for Frequent Birth and Death of MIRNA Genes. *PLoS ONE*, 2:e219.

[Fahlgren et al., 2006] Fahlgren, N., Montgomery, T. A., Howell, M. D., Allen, E., Dvorak, S. K., Alexander, A. L., and Carrington, J. C. (2006). Regulation of AUXIN RESPONSE FACTOR3 by TAS3 ta-siRNA affects developmental timing and patterning in Arabidopsis. *Curr Biol*, 16(9):939–944.

[Fattash et al., 2007] Fattash, I., Voss, B., Reski, R., Hess, W. R., and Frank, W. (2007). Evidence for the rapid expansion of microRNA-mediated regulation in early land plant evolution. *BMC Plant Biol*, 7:13.

[Fei et al., 2004] Fei, Z., Tang, X., Alba, R. M., White, J. A., Ronning, C. M., Martin, G. B., Tanksley, S. D., and Giovannoni, J. J. (2004). Comprehensive EST analysis of tomato and comparative genomics of fruit ripening. *Plant J*, 40(1):47–59.

[Fire et al., 1998] Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–811.

[Friedländer et al., 2008] Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol*, 26(4):407–415.

[Förstemann et al., 2005] Förstemann, K., Tomari, Y., Du, T., Vagin, V. V., Denli, A. M., Bratu, D. P., Klattenhoff, C., Theurkauf, W. E., and Zamore, P. D. (2005). Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol*, 3(7):e236.

[Gardner and Giegerich, 2004] Gardner, P. P. and Giegerich, R. (2004). A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5:140.

[German et al., 2008] German, M. A., Pillay, M., Jeong, D.-H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. A., Nobuta, K., German, R., Paoli, E. D., Lu, C., Schroth, G., Meyers, B. C., and Green, P. J. (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. *Nat Biotechnol*, 26(8):941–946.

[GFF, 2000] GFF (2000). GFF (General Feature Format) Specifications Document. http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml.

[Giovannoni, 2004] Giovannoni, J. J. (2004). Genetic regulation of fruit development and ripening. *Plant Cell*, 16 Suppl:S170–S180.

[Gregory et al., 2004] Gregory, R. I., Yan, K.-P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. *Nature*, 432(7014):235–240.

[Grey et al., 2005] Grey, F., Antoniewicz, A., Allen, E., Saugstad, J., McShea, A., Carrington, J. C., and Nelson, J. (2005). Identification and characterization of human cytomegalovirus-encoded microRNAs. *J Virol*, 79(18):12095–12099.

[Griffiths-Jones et al., 2006] Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., and Enright, A. J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue):D140–D144.

[Griffiths-Jones et al., 2005] Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33(Database issue):D121–D124.

[Griffiths-Jones et al., 2008] Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J. (2008). miRBase: tools for microRNA genomics. *Nucleic Acids Res*, 36(Database issue):D154–D158.

[Grundhoff et al., 2006] Grundhoff, A., Sullivan, C. S., and Ganem, D. (2006). A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. *RNA*, 12(5):733–750.

[Gunawardane et al., 2007] Gunawardane, L. S., Saito, K., Nishida, K. M., Miyoshi, K., Kawamura, Y., Nagami, T., Siomi, H., and Siomi, M. C. (2007). A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. *Science*, 315(5818):1587–1590.

[Gustafson et al., 2005] Gustafson, A. M., Allen, E., Givan, S., Smith, D., Carrington, J. C., and Kasschau, K. D. (2005). ASRP: the Arabidopsis Small RNA Project Database. *Nucleic Acids Res*, 33(Database issue):D637–D640.

[Hamilton and Baulcombe, 1999] Hamilton, A. J. and Baulcombe, D. C. (1999). A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286(5441):950–952.

[Han et al., 2004] Han, M.-H., Goud, S., Song, L., and Fedoroff, N. (2004). The Arabidopsis double-stranded RNA-binding protein HYL1 plays a role in microRNA-mediated gene regulation. *Proc Natl Acad Sci U S A*, 101(4):1093–1098.

[Harper et al., 2002] Harper, G., Hull, R., Lockhart, B., and Olszewski, N. (2002). Viral sequences integrated into plant genomes. *Annu Rev Phytopathol*, 40:119–136.

[Hartig et al., 2007] Hartig, J. V., Tomari, Y., and Förstemann, K. (2007). piRNAs–the ancient hunters of genome invaders. *Genes Dev*, 21(14):1707–1713.

[Hayes et al., 2006] Hayes, G. D., Frand, A. R., and Ruvkun, G. (2006). The mir-84 and let-7 paralogous microRNA genes of Caenorhabditis elegans direct the cessation of molting via the conserved nuclear hormone receptors NHR-23 and NHR-25. *Development*, 133(23):4631–4641.

[He et al., 2005] He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S. W., Hannon, G. J., and Hammond, S. M. (2005). A microRNA polycistron as a potential human oncogene. *Nature*, 435(7043):828–833.

[Henz et al., 2007] Henz, S. R., Cumbie, J. S., Kasschau, K. D., Lohmann, J. U., Carrington, J. C., Weigel, D., and Schmid, M. (2007). Distinct expression patterns of natural antisense transcripts in Arabidopsis. *Plant Physiol*, 144(3):1247–1255.

[Hertel et al., 2006] Hertel, J., Lindemeyer, M., Missal, K., Fried, C., Tanzer, A., Flamm, C., Hofacker, I. L., Stadler, P. F., of Bioinformatics Computer Labs 2004, S., and 2005 (2006). The expansion of the metazoan microRNA repertoire. *BMC Genomics*, 7:25.

[Hofacker, 2003] Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13):3429–3431.

[Hong and Tucker, 1998] Hong, S. B. and Tucker, M. L. (1998). Genomic organization of six tomato polygalacturonases and 5’ upstream sequence identity with tap1 and win2 genes. *Mol Gen Genet*, 258(5):479–487.

[Howell et al., 2007] Howell, M. D., Fahlgren, N., Chapman, E. J., Cumbie, J. S., Sullivan, C. M., Givan, S. A., Kasschau, K. D., and Carrington, J. C. (2007). Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell*, 19(3):926–942.

[Itaya et al., 2008] Itaya, A., Bundschuh, R., Archual, A. J., Joung, J.-G., Fei, Z., Dai, X., Zhao, P. X., Tang, Y., Nelson, R. S., and Ding, B. (2008). Small RNAs in tomato fruit and leaf development. *Biochim Biophys Acta*, 1779(2):99–107.

[Jin et al., 2008] Jin, W., Li, N., Zhang, B., Wu, F., Li, W., Guo, A., and Deng, Z. (2008). Identification and verification of microRNA in wheat (Triticum aestivum). *J Plant Res*, 121(3):351–355.

[John et al., 2004] John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C., and Marks, D. S. (2004). Human MicroRNA targets. *PLoS Biol*, 2(11):e363.

[Johnson et al., 2007] Johnson, C., Bowman, L., Adai, A. T., Vance, V., and Sundaresan, V. (2007). CSRDB: a small RNA integrated database and browser resource for cereals. *Nucleic Acids Res*, 35(Database issue):D829–D833.

[Jones-Rhoades and Bartel, 2004] Jones-Rhoades, M. W. and Bartel, D. P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell*, 14(6):787–799.

[Jones-Rhoades et al., 2006] Jones-Rhoades, M. W., Bartel, D. P., and Bartel, B. (2006). MicroRNAS and their regulatory roles in plants. *Annu Rev Plant Biol*, 57:19–53.

[Kasprzyk et al., 2004] Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E. (2004). EnsMart: a generic system for fast and flexible access to biological data. *Genome Res*, 14(1):160–169.

[Kasschau and Carrington, 1998] Kasschau, K. D. and Carrington, J. C. (1998). A counterdefensive strategy of plant viruses: suppression of posttranscriptional gene silencing. *Cell*, 95(4):461–470.

[Kasschau et al., 2007] Kasschau, K. D., Fahlgren, N., Chapman, E. J., Sullivan, C. M., Cumbie, J. S., Givan, S. A., and Carrington, J. C. (2007). Genome-Wide Profiling and Analysis of Arabidopsis siRNAs. *PLoS Biol*, 5(3):e57.

[Katiyar-Agarwal et al., 2006] Katiyar-Agarwal, S., Morgan, R., Dahlbeck, D., Borsani, O., Villegas, A., Zhu, J.-K., Staskawicz, B. J., and Jin, H. (2006). A pathogen-inducible endogenous siRNA in plant immunity. *Proc Natl Acad Sci U S A*, 103(47):18002–18007.

[Kawaji and Hayashizaki, 2008] Kawaji, H. and Hayashizaki, Y. (2008). Exploration of small RNAs. *PLoS Genet*, 4(1):e22.

[Kertesz et al., 2007] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. (2007). The role of site accessibility in microRNA target recognition. *Nat Genet*, 39(10):1278–1284.

[Kidner and Martienssen, 2005] Kidner, C. A. and Martienssen, R. A. (2005). The developmental role of microRNA in plants. *Curr Opin Plant Biol*, 8(1):38–44.

[Kim et al., 2004] Kim, J. T., Gewehr, J. E., and Martinetz, T. (2004). Binding matrix: a novel approach for binding site recognition. *J Bioinform Comput Biol*, 2(2):289–307.

[Kim et al., 2006] Kim, S.-K., Nam, J.-W., Rhee, J.-K., Lee, W.-J., and Zhang, B.-T. (2006). miTarget: microRNA target gene prediction using a support vector machine. *BMC Bioinformatics*, 7:411.

[Kim, 2005a] Kim, V. N. (2005a). MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol*, 6(5):376–385.

[Kim, 2005b] Kim, V. N. (2005b). Small RNAs: classification, biogenesis, and function. *Mol Cells*, 19(1):1–15.

[Klein and Eddy, 2003] Klein, R. J. and Eddy, S. R. (2003). RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4:44.

[Kloosterman and Plasterk, 2006] Kloosterman, W. P. and Plasterk, R. H. A. (2006). The diverse functions of microRNAs in animal development and disease. *Dev Cell*, 11(4):441–450.

[Krek et al., 2005] Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., and Rajewsky, N. (2005). Combinatorial microRNA target predictions. *Nat Genet*, 37(5):495–500.

[Krüger and Rehmsmeier, 2006] Krüger, J. and Rehmsmeier, M. (2006). RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res*, 34(Web Server issue):W451–W454.

[Kulikova et al., 2007] Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P., Castro, M.,

Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Hoad, G., Kanz, C., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Nardone, F., Pastor, M. P. G., Plaister, S., Sobhany, S., Stoehr, P., Vaughan, R., Wu, D., Zhu, W., and Apweiler, R. (2007). EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res*, 35(Database issue):D16–D20.

[Lagos-Quintana et al., 2001] 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.

[Lagos-Quintana et al., 2003] Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. (2003). New microRNAs from mouse and human. *RNA*, 9(2):175–179.

[Lagos-Quintana et al., 2002] Lagos-Quintana, M., Rauhut, R., Yalcin, A., Meyer, J., Lendeckel, W., and Tuschl, T. (2002). Identification of tissue-specific microRNAs from mouse. *Curr Biol*, 12(9):735–739.

[Lai et al., 2003] Lai, E. C., Tomancak, P., Williams, R. W., and Rubin, G. M. (2003). Computational identification of Drosophila microRNA genes. *Genome Biol*, 4(7):R42.

[Lau et al., 2001] Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294(5543):858–862.

[Lee and Ambros, 2001] Lee, R. C. and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294(5543):862–864.

[Lee et al., 1993] 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.

[Legendre et al., 2005] Legendre, M., Lambert, A., and Gautheret, D. (2005). Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, 21(7):841–845.

[Lewis et al., 2005] Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20.

[Lim et al., 2003] 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 Dev*, 17(8):991–1008.

[Lippman and Martienssen, 2004] Lippman, Z. and Martienssen, R. (2004). The role of RNA interference in heterochromatic silencing. *Nature*, 431(7006):364–370.

[Liu et al., 2004] Liu, J., He, Y., Amasino, R., and Chen, X. (2004). siRNAs targeting an intronic transposon in the regulation of natural flowering behavior in Arabidopsis. *Genes Dev*, 18(23):2873–2878.

[Llave et al., 2002] Llave, C., Xie, Z., Kasschau, K. D., and Carrington, J. C. (2002). Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. *Science*, 297(5589):2053–2056.

[Loong and Mishra, 2007a] Loong, S. N. K. and Mishra, S. K. (2007a). De Novo SVM Classification of Precursor MicroRNAs from Genomic Pseudo Hairpins Using Global and Intrinsic Folding Measures. *Bioinformatics*.

[Loong and Mishra, 2007b] Loong, S. N. K. and Mishra, S. K. (2007b). Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification. *RNA*, 13(2):170–187.

[Lowe, 2004] Lowe, T. (2004). The Arabidopsis tRNA database. `http://lowelab.ucsc.edu/GtRNAdb/Athal/`.

[Lu et al., 2008] Lu, C., Jeong, D.-H., Kulkarni, K., Pillay, M., Nobuta, K., German, R., Thatcher, S. R., Maher, C., Zhang, L., Ware, D., Liu, B., Cao, X., Meyers, B. C., and Green, P. J. (2008). Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs). *Proc Natl Acad Sci U S A*, 105(12):4951–4956.

[Lu et al., 2006] Lu, C., Kulkarni, K., Souret, F. F., MuthuValliappan, R., Tej, S. S., Poethig, R. S., Henderson, I. R., Jacobsen, S. E., Wang, W., Green, P. J., and Meyers, B. C. (2006). MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Res*, 16(10):1276–1288.

[Lu et al., 2005a] Lu, C., Tej, S. S., Luo, S., Haudenschild, C. D., Meyers, B. C., and Green, P. J. (2005a). Elucidation of the small RNA component of the transcriptome. *Science*, 309(5740):1567–1569.

[Lu et al., 2005b] Lu, S., Sun, Y.-H., Shi, R., Clark, C., Li, L., and Chiang, V. L. (2005b). Novel and mechanical stress-responsive MicroRNAs in Populus trichocarpa that are absent from Arabidopsis. *Plant Cell*, 17(8):2186–2203.

[Madany Mamlouk et al., 2003] Madany Mamlouk, A., Kim, J. T., Barth, E., Brauckmann, M., and Martinetz, T. (2003). One-Class Classification with Subgaussians. In Michaelis, B. and Krell, G., editors, *DAGM Symposium*, pages 346–353, Berlin Heidelberg. Springer Verlag.

[Manning et al., 2006] Manning, K., Tör, M., Poole, M., Hong, Y., Thompson, A. J., King, G. J., Giovannoni, J. J., and Seymour, G. B. (2006). A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. *Nat Genet*, 38(8):948–952.

[Mao et al., 2001] Mao, L., Begum, D., Goff, S. A., and Wing, R. A. (2001). Sequence and analysis of the tomato JOINTLESS locus. *Plant Physiol*, 126(3):1331–1340.

[Margulies et al., 2005] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y.-J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J.-B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu,

P., Begley, R. F., and Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380.

[Mazière and Enright, 2007] Mazière, P. and Enright, A. J. (2007). Prediction of microRNA targets. *Drug Discov Today*, 12(11-12):452–458.

[McCaskill, 1990] McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119.

[Meissner et al., 1997] Meissner, R., Jacobson, Y., Melamed, S., Levyatuv, S., Shalev, G., Ashri, A., Elkind, Y., and Levy, A. (1997). A new model system for tomato genetics. *The Plant Journal*, 12(8):1465–1472.

[Mi et al., 2008] Mi, S., Cai, T., Hu, Y., Chen, Y., Hodges, E., Ni, F., Wu, L., Li, S., Zhou, H., Long, C., Chen, S., Hannon, G. J., and Qi, Y. (2008). Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide. *Cell*, 133(1):116–127.

[Millar and Waterhouse, 2005] Millar, A. A. and Waterhouse, P. M. (2005). Plant and animal microRNAs: similarities and differences. *Funct Integr Genomics*, 5(3):129–135.

[Miranda et al., 2006] Miranda, K. C., Huynh, T., Tay, Y., Ang, Y.-S., Tam, W.-L., Thomson, A. M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell*, 126(6):1203–1217.

[Mlotshwa et al., 2008] Mlotshwa, S., Pruss, G. J., Peragine, A., Endres, M. W., Li, J., Chen, X., Poethig, R. S., Bowman, L. H., and Vance, V. (2008). DICER-LIKE2 plays a primary role in transitive silencing of transgenes in Arabidopsis. *PLoS ONE*, 3(3):e1755.

[Moissiard et al., 2007] Moissiard, G., Parizotto, E. A., Himber, C., and Voinnet, O. (2007). Transitivity in Arabidopsis can be primed, requires the redundant action of the antiviral Dicer-like 4 and Dicer-like 2, and is compromised by viral-encoded suppressor proteins. *RNA*, 13(8):1268–1278.

[Molnár et al., 2007] Molnár, A., Schwach, F., Studholme, D. J., Thuenemann, E. C., and Baulcombe, D. C. (2007). miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature*, 447(7148):1126–1129.

[Morin et al., 2008] Morin, R. D., Aksay, G., Dolgosheina, E., Ebhardt, H. A., Magrini, V., Mardis, E. R., Sahinalp, S. C., and Unrau, P. J. (2008). Comparative analysis of the small RNA transcriptomes of Pinus contorta and Oryza sativa. *Genome Res*, 18(4):571–584.

[Mosher et al., 2008] Mosher, R. A., Schwach, F., Studholme, D., and Baulcombe, D. C. (2008). PolIVb influences RNA-directed DNA methylation independently of its role in siRNA biogenesis. *Proc Natl Acad Sci U S A*, 105(8):3145–3150.

[Mueller et al., 2005] Mueller, L. A., Solow, T. H., Taylor, N., Skwarecki, B., Buels, R., Binns, J., Lin, C., Wright, M. H., Ahrens, R., Wang, Y., Herbst, E. V., Keyder, E. R., Menda, N., Zamir, D., and Tanksley, S. D. (2005). The SOL Genomics Network: a comparative resource for Solanaceae biology and beyond. *Plant Physiol*, 138(3):1310–1317.

[Noble, 2006] Noble, W. S. (2006). What is a support vector machine? *Nat Biotechnol*, 24(12):1565–1567.

[Nogueira et al., 2007] Nogueira, F. T. S., Madi, S., Chitwood, D. H., Juarez, M. T., and Timmermans, M. C. P. (2007). Two small regulatory RNAs establish opposing fates of a developmental axis. *Genes Dev*, 21(7):750–755.

[Palatnik et al., 2003] Palatnik, J. F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J. C., and Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. *Nature*, 425(6955):257–263.

[Pall et al., 2007] Pall, G. S., Codony-Servat, C., Byrne, J., Ritchie, L., and Hamilton, A. (2007). Carbodiimide-mediated cross-linking of RNA to nylon membranes improves the detection of siRNA, miRNA and piRNA by northern blot. *Nucleic Acids Res*, 35(8):e60.

[Park et al., 2002] Park, W., Li, J., Song, R., Messing, J., and Chen, X. (2002). CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana. *Curr Biol*, 12(17):1484–1495.

[Pasquinelli, 2002] Pasquinelli, A. E. (2002). MicroRNAs: deviants no longer. *Trends Genet*, 18(4):171–173.

[Pasquinelli et al., 2000] Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Müller, P., Spring, J., Srinivasan, A., Fishman, M., Finnerty, J., Corbo, J., Levine, M., Leahy, P., Davidson, E., and Ruvkun, G. (2000). Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89.

[Pearson, 2000] Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol*, 132:185–219.

[Pearson and Lipman, 1988] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–2448.

[Peragine et al., 2004] Peragine, A., Yoshikawa, M., Wu, G., Albrecht, H. L., and Poethig, R. S. (2004). SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of trans-acting siRNAs in Arabidopsis. *Genes Dev*, 18(19):2368–2379.

[Pfeffer et al., 2004] Pfeffer, S., Zavolan, M., Grässer, F. A., Chien, M., Russo, J. J., Ju, J., John, B., Enright, A. J., Marks, D., Sander, C., and Tuschl, T. (2004). Identification of virus-encoded microRNAs. *Science*, 304(5671):734–736.

[Pilcher et al., 2007] Pilcher, R. L. R., Moxon, S., Pakseresht, N., Moulton, V., Manning, K., Seymour, G., and Dalmay, T. (2007). Identification of novel small RNAs in tomato (Solanum lycopersicum). *Planta*, 226(3):709–717.

[Piriyapongsa and Jordan, 2007] Piriyapongsa, J. and Jordan, I. K. (2007). A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE*, 2(2):e203.

[Piriyapongsa et al., 2007] Piriyapongsa, J., Mariño-Ramírez, L., and Jordan, I. K. (2007). Origin and evolution of human microRNAs from transposable elements. *Genetics*, 176(2):1323–1337.

[Pontes et al., 2006] Pontes, O., Li, C. F., Nunes, P. C., Haag, J., Ream, T., Vitins, A., Jacobsen, S. E., and Pikaard, C. S. (2006). The Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center. *Cell*, 126(1):79–92.

[Prüfer et al., 2008] Prüfer, K., Stenzel, U., Dannemann, M., Green, R. E., Lachmann, M., and Kelso, J. (2008). PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics*, 24(13):1530–1531.

[Qi et al., 2006] Qi, Y., He, X., Wang, X.-J., Kohany, O., Jurka, J., and Hannon, G. J. (2006). Distinct catalytic and non-catalytic roles of ARGONAUTE4 in RNA-directed DNA methylation. *Nature*, 443(7114):1008–1012.

[R Development Core Team, 2004] R Development Core Team (2004). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[Rajagopalan et al., 2006] Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D. P. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev*, 20(24):3407–3425.

[Rajewsky, 2006] Rajewsky, N. (2006). microRNA target predictions in animals. *Nat Genet*, 38 Suppl:S8–13.

[Rathjen et al., 2006] Rathjen, T., Nicol, C., McConkey, G., and Dalmay, T. (2006). Analysis of short RNAs in the malaria parasite and its red blood cell host. *FEBS Lett*, 580(22):5185–5188.

[Reinartz et al., 2002] Reinartz, J., Bruyns, E., Lin, J.-Z., Burcham, T., Brenner, S., Bowen, B., Kramer, M., and Woychik, R. (2002). Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms. *Brief Funct Genomic Proteomic*, 1(1):95–104.

[Reinhart et al., 2000] Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–906.

[Reinhart et al., 2002] Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., and Bartel, D. P. (2002). MicroRNAs in plants. *Genes Dev*, 16(13):1616–1626.

[Rhee et al., 2003] Rhee, S. Y., Beavis, W., Berardini, T. Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L. A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D. C., Wu, Y., Xu, I., Yoo, D., Yoon, J., and Zhang, P. (2003). The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res*, 31(1):224–228.

[Rhoades et al., 2002] Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. *Cell*, 110(4):513–520.

[Schauer et al., 2002] Schauer, S. E., Jacobsen, S. E., Meinke, D. W., and Ray, A. (2002). DICER-LIKE1: blind men and elephants in Arabidopsis development. *Trends Plant Sci*, 7(11):487–491.

[Schwab et al., 2005] Schwab, R., Palatnik, J. F., Riester, M., Schommer, C., Schmid, M., and Weigel, D. (2005). Specific effects of microRNAs on the plant transcriptome. *Dev Cell*, 8(4):517–527.

[Seitz et al., 2004] Seitz, H., Royo, H., Bortolin, M.-L., Lin, S.-P., Ferguson-Smith, A. C., and Cavaillé, J. (2004). A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. *Genome Res*, 14(9):1741–1748.

[Sethupathy et al., 2006] Sethupathy, P., Corda, B., and Hatzigeorgiou, A. G. (2006). TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12(2):192–197.

[Sijen et al., 2007] Sijen, T., Steiner, F. A., Thijssen, K. L., and Plasterk, R. H. A. (2007). Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science*, 315(5809):244–247.

[Slotkin et al., 2005] Slotkin, R. K., Freeling, M., and Lisch, D. (2005). Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. *Nat Genet*, 37(6):641–644.

[Slotkin and Martienssen, 2007] Slotkin, R. K. and Martienssen, R. (2007). Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet*, 8(4):272–285.

[SQUID, 2002] SQUID (2002). SQUID package. `ftp://selab.janelia.org/pub/software/squid/`.

[Staginnus et al., 2007] Staginnus, C., Gregor, W., Mette, M. F., Teo, C. H., Borroto-Fernández, E. G., da Câmara Machado, M. L., Matzke, M., and Schwarzacher, T.

(2007). Endogenous pararetroviral sequences in tomato (Solanum lycopersicum) and related species. *BMC Plant Biol*, 7:24.

[Stormo, 2000] Stormo, G. D. (2000). DNA Binding Sites: Representation and Discovery. *Bioinformatics*, 16:16–23.

[Subramanian et al., 2008] Subramanian, S., Fu, Y., Sunkar, R., Barbazuk, W. B., Zhu, J.-K., and Yu, O. (2008). Novel and nodulation-regulated microRNAs in soybean roots. *BMC Genomics*, 9:160.

[Sunkar and Jagadeeswaran, 2008] Sunkar, R. and Jagadeeswaran, G. (2008). In silico identification of conserved microRNAs in large number of diverse plant species. *BMC Plant Biol*, 8:37.

[Sunkar and Zhu, 2004] Sunkar, R. and Zhu, J.-K. (2004). Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell*, 16(8):2001–2019.

[Swarbreck et al., 2008] Swarbreck, D., Wilks, C., Lamesch, P., Berardini, T. Z., Garcia-Hernandez, M., Foerster, H., Li, D., Meyer, T., Muller, R., Ploetz, L., Radenbaugh, A., Singh, S., Swing, V., Tissier, C., Zhang, P., and Huala, E. (2008). The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res*, 36(Database issue):D1009–D1014.

[Tanzer and Stadler, 2004] Tanzer, A. and Stadler, P. F. (2004). Molecular evolution of a microRNA cluster. *J Mol Biol*, 339(2):327–335.

[Turner et al., 1987] Turner, D. H., Sugimoto, N., Jaeger, J. A., Longfellow, C. E., Freier, S. M., and Kierzek, R. (1987). Improved parameters for prediction of RNA structure. *Cold Spring Harb Symp Quant Biol*, 52:123–133.

[Vapnik, 1998] Vapnik, V. (1998). *Statistical Learning Theory*. Wiley: New York.

[Vaucheret, 2006] Vaucheret, H. (2006). Post-transcriptional small RNA pathways in plants: mechanisms and regulations. *Genes Dev*, 20(7):759–771.

[Vaucheret et al., 2004] Vaucheret, H., Vazquez, F., Crété, P., and Bartel, D. P. (2004). The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development. *Genes Dev*, 18(10):1187–1197.

[Vaughn et al., 2007] Vaughn, M. W., Ic, M. T., Lippman, Z., Jiang, H., Carrasquillo, R., Rabinowicz, P. D., Dedhia, N., McCombie, W. R., Agier, N., Bulski, A., Colot, V., Doerge, R. W., and Martienssen, R. A. (2007). Epigenetic Natural Variation in Arabidopsis thaliana. *PLoS Biol*, 5(7):e174.

[Vazquez et al., 2004] Vazquez, F., Vaucheret, H., Rajagopalan, R., Lepers, C., Gasciolli, V., Mallory, A. C., Hilbert, J.-L., Bartel, D. P., and Crété, P. (2004). Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs. *Mol Cell*, 16(1):69–79.

[Voinnet, 2005] Voinnet, O. (2005). Non-cell autonomous RNA silencing. *FEBS Lett*, 579(26):5858–5871.

[Voinnet et al., 1998] Voinnet, O., Vain, P., Angell, S., and Baulcombe, D. C. (1998). Systemic spread of sequence-specific transgene RNA degradation in plants is initiated by localized introduction of ectopic promoterless DNA. *Cell*, 95(2):177–187.

[von Besser et al., 2006] von Besser, K., Frank, A. C., Johnson, M. A., and Preuss, D. (2006). Arabidopsis HAP2 (GCS1) is a sperm-specific gene required for pollen tube guidance and fertilization. *Development*, 133(23):4761–4769.

[Vrebalov et al., 2002] Vrebalov, J., Ruezinsky, D., Padmanabhan, V., White, R., Medrano, D., Drake, R., Schuch, W., and Giovannoni, J. (2002). A MADS-box gene necessary for fruit ripening at the tomato ripening-inhibitor (rin) locus. *Science*, 296(5566):343–346.

[Válóczi et al., 2006] Válóczi, A., Várallyay, E., Kauppinen, S., Burgyán, J., and Havelda, Z. (2006). Spatio-temporal accumulation of microRNAs is highly coordinated in developing plant tissues. *Plant J*, 47(1):140–151.

[Wang et al., 2007] Wang, B., Doench, J. G., and Novina, C. D. (2007). Analysis of microRNA effector functions in vitro. *Methods*, 43(2):91–104.

[Wang et al., 2004a] Wang, J.-F., Zhou, H., Chen, Y.-Q., Luo, Q.-J., and Qu, L.-H. (2004a). Identification of 20 microRNAs from Oryza sativa. *Nucleic Acids Res*, 32(5):1688–1695.

[Wang and Naqa, 2008] Wang, X. and Naqa, I. M. E. (2008). Prediction of both conserved and nonconserved microRNA targets in animals. *Bioinformatics*, 24(3):325–332.

[Wang et al., 2005] 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.

[Wang et al., 2004b] Wang, X.-J., Reyes, J. L., Chua, N.-H., and Gaasterland, T. (2004b). Prediction and identification of Arabidopsis thaliana microRNAs and their mRNA targets. *Genome Biol*, 5(9):R65.

[Watanabe et al., 2006] Watanabe, Y., Yachie, N., Numata, K., Saito, R., Kanai, A., and Tomita, M. (2006). Computational analysis of microRNA targets in Caenorhabditis elegans. *Gene*, 365:2–10.

[Weber, 2005] Weber, M. J. (2005). New human and mouse microRNA genes found by homology search. *FEBS J*, 272(1):59–73.

[Wenkel et al., 2006] Wenkel, S., Turck, F., Singer, K., Gissot, L., Gourrierec, J. L., Samach, A., and Coupland, G. (2006). CONSTANS and the CCAAT box binding complex share a functionally important domain and interact to regulate flowering of Arabidopsis. *Plant Cell*, 18(11):2971–2984.

[Wienholds et al., 2005] Wienholds, E., Kloosterman, W. P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., Horvitz, H. R., Kauppinen, S., and Plasterk, R. H. A. (2005). MicroRNA expression in zebrafish embryonic development. *Science*, 309(5732):310–311.

[Williams et al., 2005] Williams, L., Grigg, S. P., Xie, M., Christensen, S., and Fletcher, J. C. (2005). Regulation of Arabidopsis shoot apical meristem and lateral organ formation by microRNA miR166g and its AtHD-ZIP target genes. *Development*, 132(16):3657–3668.

[Wolff et al., 2003] Wolff, H., Brack-Werner, R., Neumann, M., Werner, T., and Schneider, R. (2003). Integrated Functional and Bioinformatics Approach for the Identification and Experimental Verification of RNA Signals: Application to HIV-1 INS. *Nucleic Acids Research*, 31:2839–2851.

[Workman and Krogh, 1999] Workman, C. and Krogh, A. (1999). No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res*, 27(24):4816–4822.

[Xie et al., 2004] Xie, Z., Johansen, L. K., Gustafson, A. M., Kasschau, K. D., Lellis, A. D., Zilberman, D., Jacobsen, S. E., and Carrington, J. C. (2004). Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol*, 2(5):E104.

[Xu et al., 2003] Xu, P., Vernooy, S. Y., Guo, M., and Hay, B. A. (2003). The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Curr Biol*, 13(9):790–795.

[Xue et al., 2005] Xue, C., Li, F., He, T., Liu, G.-P., Li, Y., and Zhang, X. (2005). Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine. *BMC Bioinformatics*, 6:310.

[Yao et al., 2007] Yao, Y., Guo, G., Ni, Z., Sunkar, R., Du, J., Zhu, J.-K., and Sun, Q. (2007). Cloning and characterization of microRNAs from wheat (Triticum aestivum L.). *Genome Biol*, 8(6):R96.

[Zhang et al., 2006a] Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P., and Anderson, T. A. (2006a). Conservation and divergence of plant microRNA genes. *Plant J*, 46(2):243–259.

[Zhang et al., 2007] Zhang, B., Wang, Q., and Pan, X. (2007). MicroRNAs and their regulatory roles in animals and plants. *J Cell Physiol*, 210(2):279–289.

[Zhang et al., 2006b] Zhang, B. H., Pan, X. P., Cox, S. B., Cobb, G. P., and Anderson, T. A. (2006b). Evidence that miRNAs are different from other RNAs. *Cell Mol Life Sci*, 63(2):246–254.

[Zhang et al., 2006c] Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W.-L., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E., and Ecker, J. R. (2006c). Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. *Cell*, 126(6):1189–1201.

[Zhang, 2005] Zhang, Y. (2005). miRU: an automated plant miRNA target prediction server. *Nucleic Acids Res*, 33(Web Server issue):W701–W704.

[Zilberman et al., 2007] Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet*, 39(1):61–69.

[Zuker, 2003] Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–3415.

[Zuker and Stiegler, 1981] Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148.

Appendix A

A.1 `miRCat` Testing Results

A.1.1 *Arabidopsis* GSM118373 results

Results from `miRCat` (using default parameters) when run on the 454 *Arabidopsis thaliana* leaf sRNA set from Rajagopalan *et al*. [Rajagopalan et al., 2006].

Column ”miR” shows the miRBase accession of the miRNA (if availalbe), column ”chr” shows the *Arabidopsis thaliana* chromosome the sequence maps to, column ”start” shows the start position of the predicted miRNA, column ”end” shows the start position of the predicted miRNA, column ”ori” shows the orientation of the miRNA (either Watson "+" or Crick "-" strand). Column ”abun” shows the abundance of the sRNA in the 454 dataset, column ”seq” shows the sRNA sequence, column ”len” shows the length of the sRNA sequence, column ”g. hits” shows the number of times this sequence maps to the reference genome, column ”h. len” shows the length of the predicted miRNA precursor structure. Column ”G/C%” shows the

percentage G/C composition of the miRNA hairpin sequence, column ”MFE” shows the minimum free energy of the folded miRNA precursor sequence, column ”AMFE” shows the MFE per 100nt (therefore normalising the MFE), column ”p-value” shows the `randfold` p-value for the predicted hairpin precursor (using 100 randomisations). Column ”miRNA*” gives the sequence of any potential miRNA* sequences present in the predicted precursor.

Table A.1: miRNAs predicted in the GSM118373 454 leaf sRNA dataset [Rajagopalan et al., 2006] by `miRCat` using default settings

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
165a	1	78932	78952	-	24	TCGGACCAGGCTTCATCCCCC	21	2	154	40.91	-57.7	-37.47	0.01	GGAATGTTGTCTGGATCGAGG
171b	1	3961367	3961387	-	233	TTGAGCCGTGCCAATATCACG	21	2	117	45.3	-46.7	-39.91	0.01	CGAGATATTAGTGCGGTTCAA
472a	1	4182141	4182162	-	5	TTTTTCCTACTCCGCCCATACC	22	1	222	40.99	-91.6	-41.26	0.01	NO
159b	1	6220806	6220826	+	640	TTTGGATTGAAGGGAGCTCTT	21	1	259	38.61	-95.16	-36.74	0.01	NO
169h	1	6695535	6695555	-	126	TAGCCAAGGATGACTTGCCTG	21	7	188	39.36	-79.1	-42.07	0.01	NO
400a	1	11785927	11785947	+	15	TATGAGAGTATTATAAGTCAC	21	2	169	34.32	-53.8	-31.83	0.01	AAGTGACTTATGATAATCTCA
773a	1	13067291	13067312	+	5	TTTGCTTCCAGCTTTTGTCTCC	22	1	198	42.42	-93	-46.97	0.01	NO
N/A	1	16006805	16006826	-	5	TTAACAAATCTGGTGTTTTACA	22	15	139	26.62	-43.5	-31.29	0.01	NO
161a	1	17829399	17829419	+	4593	TGAAAGTGACTACATCGGGGT	21	1	122	39.34	-53.5	-43.85	0.01	ACCCTGGTTTAGTCACTTTCA
169d	1	20043224	20043244	-	25	TGAGCCAAGGATGACTTGCCG	21	4	201	35.82	-73.57	-36.6	0.01	GCAAGTTGACCTTGGCTCTGT
169e	1	20045254	20045274	+	25	TGAGCCAAGGATGACTTGCCG	21	4	208	36.54	-70.81	-34.04	0.01	NO
846a	1	22581327	22581347	+	51	TTGAATTGAAGTGCTTGAATT	21	1	267	36.33	-92.1	-34.49	0.01	CATTCAAGGACTTCTATTCAG
171c	1	22933760	22933780	-	233	TTGAGCCGTGCCAATATCACG	21	2	158	37.97	-64.22	-40.65	0.01	NO
399b	1	23349052	23349072	-	27	TGCCAAAGGAGAGTTGCCCTG	21	2	162	38.27	-61.4	-37.9	0.01	NO
157a	1	24916939	24916959	-	73	TTGACAGAAGATAGAGAGCAC	21	3	166	36.14	-63.8	-38.43	0.01	GCTCTCTAGCCTTCTGTCATC
157b	1	24924767	24924787	+	73	TTGACAGAAGATAGAGAGCAC	21	3	132	44.7	-57.06	-43.23	0.01	GCTCTCTAGCCTTCTGTCATC
839a	1	25282801	25282821	-	7	TACCAACCTTTCATCGTTCCC	21	1	340	41.18	-206.1	-60.62	0.01	NO
159a	1	27716895	27716915	-	4046	TTTGGATTGAAGGGAGCTCTA	21	1	184	39.13	-75.9	-41.25	0.01	NO
840a	2	771491	771512	-	22	ACACTGAAGGACCTAAACTAAC	22	1	306	43.46	-153.39	-50.13	0.01	TTGTTTAGGTCCCTTAGTTTCT
398a	2	1041009	1041028	+	41	TGTGTTCTCAGGTCACCCCT	20	3	179	36.31	-62.7	-35.03	0.01	NO
396a	2	4149524	4149545	-	12	TTCCACAGCTTTCTTGAACTGC	22	1	165	35.15	-57.7	-34.97	0.01	NO
156a	2	10683613	10683632	-	236	TGACAGAAGAGAGTGAGCAC	20	6	103	46.6	-50	-48.54	0.01	NO
825a	2	11166852	11166872	+	134	TTCTCAAGAAGGTGCATGAAC	21	1	102	38.24	-42.1	-41.27	0.01	NO
172a	2	11949995	11950015	-	835	AGAATCTTGATGATGCTGCAT	21	2	174	46.55	-67.77	-38.95	0.01	NO
390a	2	16069101	16069121	+	39	CGCTATCCATCCTGAGTTTCA	21	1	107	42.06	-51.6	-48.22	0.01	AAGCTCAGGAGGGATAGCGCC
160a	2	16347360	16347380	+	608	TGCCTGGCTCCCTGTATGCCA	21	3	116	44.83	-54.9	-47.33	0.01	GCGTATGAGGAGCCATGCATA
N/A	2	18626203	18626223	+	1556	TGATTGAGCCGCGCCAATATC	21	4	113	54.87	-40.92	-36.21	0.05	NO
166a	2	19183311	19183331	+	221	TCGGACCAGGCTTCATTCCCC	21	7	210	43.33	-73.85	-35.17	0.01	GGACTGTTGTCTGGCTCGAGG
408a	2	19327003	19327023	+	342	ATGCACTGCCTCTTCCCTGGC	21	1	159	41.51	-50.7	-31.89	0.01	CAGGGAACAAGCAGAGCATGG
403a	2	19422147	19422168	+	1070	TGTTTTGTGCTTGAATCTAATT	22	1	128	32.81	-42.29	-33.04	0.01	NO
164a	2	19527840	19527860	+	1015	TGGAGAAGCAGGGCACGTGCA	21	2	113	46.9	-48.2	-42.65	0.01	NO
167c	3	1306756	1306776	-	10	TAAGCTGCCAGCATGATCTTG	21	1	159	34.59	-66.9	-42.08	0.01	NO
158a	3	3366354	3366373	-	106	TCCCAAATGTAGACAAAGCA	20	1	126	40.48	-43.3	-34.37	0.01	CTTTGTCTACAATTTTGGAA
172c	3	3599797	3599817	-	18	AGAATCTTGATGATGCTGCAG	21	2	176	36.93	-63.56	-36.11	0.01	NO
823a	3	4496833	4496853	-	14	TGGGTGGTGATCATATAAGAT	21	1	97	32.99	-42.3	-43.61	0.01	NO
169f	3	4805806	4805826	-	25	TGAGCCAAGGATGACTTGCCG	21	4	195	42.05	-80.2	-41.13	0.01	GCAAGTTGACCTTGGCTCTGC
167a	3	8108097	8108117	+	14850	TGAAGCTGCCAGCATGATCTA	21	2	136	45.59	-57.1	-41.99	0.01	TAGATCATGTTCGCAGTTTCA
173a	3	8236168	8236189	+	15	TTCGCTTGCAGAGAGAAATCAC	22	1	158	36.08	-51.56	-32.63	0.01	NO
169i	3	9873343	9873363	-	126	TAGCCAAGGATGACTTGCCTG	21	7	206	35.92	-88.5	-42.96	0.01	NO

Continued on Next Page...

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
169j	3	9873720	9873740	-	126	TAGCCAAGGATGACTTGCCTG	21	7	182	35.16	-80.44	-44.2	0.01	NO
169k	3	9876912	9876932	-	126	TAGCCAAGGATGACTTGCCTG	21	7	206	36.41	-84.9	-41.21	0.01	NO
169l	3	9877277	9877297	-	126	TAGCCAAGGATGACTTGCCTG	21	7	149	36.24	-62.54	-41.97	0.01	NO
169m	3	9879555	9879575	-	126	TAGCCAAGGATGACTTGCCTG	21	7	208	37.5	-91.81	-44.14	0.01	NO
169n	3	9879927	9879947	-	126	TAGCCAAGGATGACTTGCCTG	21	7	295	35.93	-113.1	-38.34	0.01	NO
171a	3	19084500	19084520	+	1556	TGATTGAGCCGCGCCAATATC	21	4	123	43.09	-46.6	-37.89	0.01	TATTGGCCTGGTTCACTCAGA
172d	3	20598970	20598990	+	18	AGAATCTTGATGATGCTGCAG	21	2	143	35.66	-51.2	-35.8	0.01	GCAACATCTTCAAGATTCAGA
827a	3	22133788	22133808	-	8	TTAGATGACCATCAACAAACT	21	1	131	38.17	-38.83	-29.64	0.01	NO
166b	3	22933276	22933296	+	221	TCGGACCAGGCTTCATTCCCC	21	7	147	42.86	-62.16	-42.29	0.01	GGACTGTTGTCTGGCTCGAGG
167b	3	23417152	23417172	+	14850	TGAAGCTGCCAGCATGATCTA	21	2	142	44.37	-55.3	-38.94	0.01	NO
165b	4	369856	369876	-	24	TCGGACCAGGCTTCATCCCCC	21	2	182	37.91	-65.3	-35.88	0.01	NO
N/A	4	2179149	2179169	+	6	ACAATGCCGAATCTTGAACAA	21	5	298	34.9	-116.1	-38.96	0.01	NO
397a	4	2625958	2625978	+	27	TCATTGAGTGCAGCGTTGATG	21	1	128	32.81	-51.4	-40.16	0.01	NO
N/A	4	3163240	3163260	+	9	TGGCTGGGTGATGAAGTAAGT	21	2	102	38.24	-26.7	-26.18	0.03	NO
850a	4	7845752	7845773	+	5	TAAGATCCGGACTACAACAAAG	22	1	465	35.05	-165.6	-35.61	0.01	NO
397b	4	7878726	7878746	-	107	TCATTGAGTGCATCGTTGATG	21	1	126	28.57	-40.1	-31.83	0.01	NO
160b	4	9888999	9889019	+	608	TGCCTGGCTCCCTGTATGCCA	21	3	89	46.07	-42.3	-47.53	0.01	NO
168a	4	10578663	10578683	+	3364	TCGCTTGGTGCAGGTCGGGAA	21	2	155	50.97	-71.4	-46.06	0.01	GATCCCGCCTTGCATCAACTG
N/A	4	11224199	11224219	-	8	TGATGTGTCATTTATAGGGAG	21	1	95	25.26	-30.6	-32.21	0.01	NO
169g	4	11483106	11483126	-	25	TGAGCCAAGGATGACTTGCCG	21	4	161	42.86	-72.3	-44.91	0.01	GCAAGTTGACCTTGGCTCTGT
319a	4	12353119	12353139	+	9	TTGGACTGAAGGGAGCTCCCT	21	2	203	39.9	-83.09	-40.93	0.01	NO
156b	4	15074951	15074970	+	236	TGACAGAAGAGAGTGAGCAC	20	6	182	47.8	-93.2	-51.21	0.01	NO
156c	4	15415497	15415516	-	236	TGACAGAAGAGAGTGAGCAC	20	6	85	48.24	-44.5	-52.35	0.01	NO
164b	5	287583	287603	+	1015	TGGAGAAGCAGGGCACGTGCA	21	2	159	41.51	-67.4	-42.39	0.01	CATGTGCCCATCTTCACCATC
822a	5	897286	897306	-	26	TGCGGGAAGCATTTGCACATG	21	1	337	32.64	-177.6	-52.7	0.01	ACATGTGCAAATGCTTTCTAC
172b	5	1188212	1188232	-	835	AGAATCTTGATGATGCTGCAT	21	2	117	38.46	-48.9	-41.79	0.01	GCAGCACCATTAAGATTCACA
162a	5	2634937	2634957	-	70	TCGATAAACCTCTGCATCCAG	21	2	125	44	-53.3	-42.64	0.01	NO
166c	5	2838738	2838758	+	221	TCGGACCAGGCTTCATTCCCC	21	7	139	42.45	-46.5	-33.45	0.01	GGATTGTTGTCTGGCTCGAGG
166d	5	2840709	2840729	+	221	TCGGACCAGGCTTCATTCCCC	21	7	113	44.25	-41.6	-36.81	0.01	NO
156d	5	3456714	3456733	-	236	TGACAGAAGAGAGTGAGCAC	20	6	115	40	-56.4	-49.04	0.01	NO
156e	5	3867214	3867233	+	236	TGACAGAAGAGAGTGAGCAC	20	6	202	42.57	-80.9	-40.05	0.01	NO
848a	5	4479450	4479471	-	22	TGACATGGGACTGCCTAAGCTA	22	1	179	39.66	-67.1	-37.49	0.01	NO
398b	5	4691110	4691130	+	2267	TGTGTTCTCAGGTCACCCCTG	21	2	135	48.15	-59.9	-44.37	0.01	AGGGTTGATATGAGAACACAC
398c	5	4694781	4694801	+	2267	TGTGTTCTCAGGTCACCCCTG	21	2	157	44.59	-57.7	-36.75	0.01	AGGGTTGATATGAGAACACAC
162b	5	7740613	7740633	-	70	TCGATAAACCTCTGCATCCAG	21	2	115	42.61	-50.3	-43.74	0.01	NO
169b	5	8527595	8527615	+	7	GGCAAGTTGTCCTTCGGCTAC	21	1	181	40.88	-82.3	-45.47	0.01	TGCAGCCAAGGATGACTTGCC
156f	5	9136129	9136148	+	236	TGACAGAAGAGAGTGAGCAC	20	6	148	50	-66.82	-45.15	0.01	NO
164c	5	9852688	9852708	+	46	TGGAGAAGCAGGGCACGTGCG	21	1	116	48.28	-51.8	-44.66	0.01	CACGTGTTCTACTACTCCAAC
N/A	5	12107291	12107310	+	8	GGCATGATTGGTTGGGTTGT	20	2	135	32.59	-41.57	-30.79	0.01	NO
396b	5	13629038	13629058	+	23	TTCCACAGCTTTCTTGAACTT	21	1	155	30.97	-55.4	-35.74	0.01	GCTCAAGAAAGCTGTGGGAAA
166e	5	16792752	16792772	-	221	TCGGACCAGGCTTCATTCCCC	21	7	143	37.06	-49.3	-34.48	0.01	GGAATGTTGTCTGGCACGAGG
166f	5	17533605	17533625	+	221	TCGGACCAGGCTTCATTCCCC	21	7	105	43.81	-42.1	-40.1	0.01	NO
168b	5	18376100	18376120	-	3364	TCGCTTGGTGCAGGTCGGGAA	21	2	156	50	-68	-43.59	0.01	CCCGTCTTGTATCAACTGAAT
160c	5	19026385	19026405	-	608	TGCCTGGCTCCCTGTATGCCA	21	3	117	47.01	-55	-47.01	0.01	GCGTACAAGGAGTCAAGCATG
390b	5	23654187	23654207	+	38	AAGCTCAGGAGGGATAGCGCC	21	2	161	40.99	-64.6	-40.12	0.01	TGGCGCTATCCATCCTGAGTT
172e	5	24005710	24005729	+	33	GCAGCACCATTAAGATTCAC	20	2	136	47.79	-62.5	-45.96	0.01	NO
391a	5	24310386	24310406	+	259	TTCGCAGGAGAGATAGCGCCA	21	1	251	36.65	-74.1	-29.52	0.01	ACGGTATCTCTCCTACGTAGC
399c	5	24979794	24979814	+	27	TGCCAAAGGAGAGTTGCCCTG	21	2	125	45.6	-66.59	-53.27	0.01	GGGCATCTTTCTATTGGCAGG
166g	5	25522108	25522128	+	221	TCGGACCAGGCTTCATTCCCC	21	7	122	43.44	-47.9	-39.26	0.01	NO
170a	5	26428820	26428840	-	279	TGATTGAGCCGTGTCAATATC	21	1	93	46.24	-40.49	-43.54	0.01	TATTGGCCTGGTTCACTCAGA

A.1.2 *Arabidopsis* combined results

Results from `miRCat` (using default parameters) when run on the 454 *Arabidopsis thaliana* leaf sRNA set from Kasschau *et al*. [Kasschau et al., 2007].

Column ”miR” shows the miRBase accession of the miRNA (if availalbe), column ”chr” shows the *Arabidopsis thaliana* chromosome the sequence maps to, column ”start” shows the start position of the predicted miRNA, column ”end” shows the start position of the predicted miRNA, column ”ori” shows the orientation of the miRNA (either Watson "+" or Crick "-" strand). Column ”abun” shows the abundance of the sRNA in the 454 dataset, column ”seq” shows the sRNA sequence, column ”len” shows the length of the sRNA sequence, column ”g. hits” shows the number of times this sequence maps to the reference genome, column ”h. len” shows the length of the predicted miRNA precursor structure. Column ”G/C%” shows the percentage G/C composition of the miRNA hairpin sequence, column ”MFE” shows the minimum free energy of the folded miRNA precursor sequence, column ”AMFE” shows the MFE per 100nt (therefore normalising the MFE), column ”p-value” shows the `randfold` p-value for the predicted hairpin precursor (using 100 randomisations). Column ”miRNA*” gives the sequence of any potential miRNA* sequences present in the predicted precursor.

Table A.2: miRNAs predicted in the GSM118373 454 leaf sRNA dataset [Kasschau et al., 2007] by `miRCat` using default settings

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
171b	1	3961430	3961450	-	8	AGATATTAGTGCGGTTCAATC	21	1	102	47.06	-44.8	-43.92	0.01	NO
159b	1	6220806	6220826	+	109	TTTGGATTGAAGGGAGCTCTT	21	1	259	38.61	-95.16	-36.74	0.01	GAGCTCCTTGAAGTTCAATGG
169h	1	6695535	6695555	-	15	TAGCCAAGGATGACTTGCCTG	21	7	188	39.36	-79.1	-42.07	0.01	NO
400a	1	11785927	11785947	+	6	TATGAGAGTATTATAAGTCAC	21	2	169	34.32	-53.8	-31.83	0.01	NO
161a	1	17829399	17829419	+	121	TGAAAGTGACTACATCGGGGT	21	1	122	39.34	-53.5	-43.85	0.01	NO

Continued on Next Page. . .

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
846a	1	22581327	22581347	+	12	TTGAATTGAAGTGCTTGAATT	21	1	267	36.33	-92.1	-34.49	0.01	NO
171c	1	22933822	22933842	-	10	AGATATTGGTGCGGTTCAATC	21	1	79	37.97	-44.6	-56.46	0.01	NO
157a	1	24916939	24916959	-	12	TTGACAGAAGATAGAGAGCAC	21	3	166	36.14	-63.8	-38.43	0.01	NO
157b	1	24924767	24924787	+	12	TTGACAGAAGATAGAGAGCAC	21	3	132	44.7	-57.06	-43.23	0.01	NO
159a	1	27716895	27716915	-	220	TTTGGATTGAAGGGAGCTCTA	21	1	184	39.13	-75.9	-41.25	0.01	NO
396a	2	4149525	4149545	-	51	TTCCACAGCTTTCTTGAACTG	21	1	165	35.15	-57.7	-34.97	0.01	GTTCAATAAAGCTGTGGGAAG
825a	2	11166852	11166872	+	43	TTCTCAAGAAGGTGCATGAAC	21	1	102	38.24	-42.1	-41.27	0.01	NO
172a	2	11949995	11950014	-	34	GAATCTTGATGATGCTGCAT	20	3	174	46.55	-67.77	-38.95	0.01	NO
160a	2	16347360	16347380	+	31	TGCCTGGCTCCCTGTATGCCA	21	3	116	44.83	-54.9	-47.33	0.01	NO
393a	2	16659189	16659210	+	5	TCCAAAGGGATCGCATTGATCC	22	2	147	36.73	-54.3	-36.94	0.01	NO
N/A	2	18626203	18626223	+	56	TGATTGAGCCGCGCCAATATC	21	4	113	54.87	-40.92	-36.21	0.03	NO
166a	2	19183311	19183331	+	42	TCGGACCAGGCTTCATTCCCC	21	7	210	43.33	-73.85	-35.17	0.01	NO
408a	2	19327003	19327023	+	6	ATGCACTGCCTCTTCCCTGGC	21	1	159	41.51	-50.7	-31.89	0.01	NO
403a	2	19422147	19422168	+	7	TGTTTTGTGCTTGAATCTAATT	22	1	128	32.81	-42.29	-33.04	0.01	NO
164a	2	19527840	19527860	+	19	TGGAGAAGCAGGGCACGTGCA	21	2	113	46.9	-48.2	-42.65	0.01	NO
158a	3	3366354	3366373	-	52	TCCCAAATGTAGACAAAGCA	20	1	126	40.48	-43.3	-34.37	0.01	NO
823a	3	4496833	4496853	-	7	TGGGTGGTGATCATATAAGAT	21	1	97	32.99	-42.3	-43.61	0.01	NO
167a	3	8108097	8108117	+	139	TGAAGCTGCCAGCATGATCTA	21	2	136	45.59	-57.1	-41.99	0.01	GATCATGTTCGCAGTTTCACC
169i	3	9873343	9873363	-	15	TAGCCAAGGATGACTTGCCTG	21	7	206	35.92	-88.5	-42.96	0.01	NO
169j	3	9873720	9873740	-	15	TAGCCAAGGATGACTTGCCTG	21	7	182	35.16	-80.44	-44.2	0.01	NO
169k	3	9876912	9876932	-	15	TAGCCAAGGATGACTTGCCTG	21	7	206	36.41	-84.9	-41.21	0.01	NO
169l	3	9877277	9877297	-	15	TAGCCAAGGATGACTTGCCTG	21	7	149	36.24	-62.54	-41.97	0.01	NO
169m	3	9879555	9879575	-	15	TAGCCAAGGATGACTTGCCTG	21	7	208	37.5	-91.81	-44.14	0.01	NO
169n	3	9879927	9879947	-	15	TAGCCAAGGATGACTTGCCTG	21	7	295	35.93	-113.1	-38.34	0.01	NO
171a	3	19084500	19084520	+	56	TGATTGAGCCGCGCCAATATC	21	4	123	43.09	-46.6	-37.89	0.01	NO
393b	3	20702636	20702657	+	5	TCCAAAGGGATCGCATTGATCC	22	2	174	35.06	-67.6	-38.85	0.01	NO
166b	3	22933276	22933296	+	42	TCGGACCAGGCTTCATTCCCC	21	7	147	42.86	-62.16	-42.29	0.01	NO
167b	3	23417152	23417172	+	139	TGAAGCTGCCAGCATGATCTA	21	2	142	44.37	-55.3	-38.94	0.01	NO
397b	4	7878726	7878746	-	12	TCATTGAGTGCATCGTTGATG	21	1	126	28.57	-40.1	-31.83	0.01	NO
N/A	4	7890503	7890523	-	11	TTGATGTTGTGCTACGATACA	21	1	298	36.91	-124.8	-41.88	0.01	NO
160b	4	9888999	9889019	+	31	TGCCTGGCTCCCTGTATGCCA	21	3	89	46.07	-42.3	-47.53	0.01	NO
168a	4	10578663	10578683	+	78	TCGCTTGGTGCAGGTCGGGAA	21	2	155	50.97	-71.4	-46.06	0.01	CCCGCCTTGCATCAACTGAAT
164b	5	287583	287603	+	19	TGGAGAAGCAGGGCACGTGCA	21	2	159	41.51	-67.4	-42.39	0.01	NO
822a	5	897045	897065	-	7	AAACAATATACGTTGCATCCC	21	1	291	32.65	-149.8	-51.48	0.01	NO
172b	5	1188212	1188231	-	34	GAATCTTGATGATGCTGCAT	20	3	117	38.46	-48.9	-41.79	0.01	NO
162a	5	2634937	2634957	-	6	TCGATAAACCTCTGCATCCAG	21	2	125	44	-53.3	-42.64	0.01	NO
166c	5	2838738	2838758	+	42	TCGGACCAGGCTTCATTCCCC	21	7	139	42.45	-46.5	-33.45	0.01	NO
166d	5	2840709	2840729	+	42	TCGGACCAGGCTTCATTCCCC	21	7	113	44.25	-41.6	-36.81	0.01	NO
162b	5	7740613	7740633	-	6	TCGATAAACCTCTGCATCCAG	21	2	115	42.61	-50.3	-43.74	0.01	NO
164c	5	9852751	9852771	+	13	CACGTGTTCTACTACTCCAAC	21	1	116	48.28	-51.8	-44.66	0.01	NO
396b	5	13629038	13629058	+	12	TTCCACAGCTTTCTTGAACTT	21	1	155	30.97	-55.4	-35.74	0.01	GCTCAAGAAAGCTGTGGGAAA
166e	5	16792752	16792772	-	42	TCGGACCAGGCTTCATTCCCC	21	7	143	37.06	-49.3	-34.48	0.01	NO
166f	5	17533605	17533625	+	42	TCGGACCAGGCTTCATTCCCC	21	7	105	43.81	-42.1	-40.1	0.01	NO
168b	5	18376100	18376120	-	78	TCGCTTGGTGCAGGTCGGGAA	21	2	156	50	-68	-43.59	0.01	NO
160c	5	19026385	19026405	-	31	TGCCTGGCTCCCTGTATGCCA	21	3	117	47.01	-55	-47.01	0.01	NO
172e	5	24005793	24005812	+	34	GAATCTTGATGATGCTGCAT	20	3	197	45.18	-72.75	-36.93	0.01	NO
391a	5	24310437	24310457	+	57	ACGGTATCTCTCCTACGTAGC	21	1	206	38.35	-64.7	-31.41	0.01	TTCGCAGGAGAGATAGCGCCA
166g	5	25522108	25522128	+	42	TCGGACCAGGCTTCATTCCCC	21	7	122	43.44	-47.9	-39.26	0.01	NO

A.1.3 *Arabidopsis* combined results

Results from `miRCat` (using default parameters) when run on the combined 454 *Arabidopsis thaliana* sets (GEO accessionsGSM118372, GSM118373, GSM149079, GSM154336, GSM154370, GSM257235, GSM118375, GSM121455 and GSM149080.

Column ”miR” shows the miRBase accession of the miRNA (if availalbe), column ”chr” shows the *Arabidopsis thaliana* chromosome the sequence maps to, column ”start” shows the start position of the predicted miRNA, column ”end” shows the start position of the predicted miRNA, column ”ori” shows the orientation of the miRNA (either Watson "+" or Crick "-" strand). Column ”abun” shows the abundance of the sRNA in the 454 dataset, column ”seq” shows the sRNA sequence, column ”len” shows the length of the sRNA sequence, column ”g. hits” shows the number of times this sequence maps to the reference genome, column ”h. len” shows the length of the predicted miRNA precursor structure. Column ”G/C%” shows the percentage G/C composition of the miRNA hairpin sequence, column ”MFE” shows the minimum free energy of the folded miRNA precursor sequence, column ”AMFE” shows the MFE per 100nt (therefore normalising the MFE), column ”p-value” shows the `randfold` p-value for the predicted hairpin precursor (using 100 randomisations). Column ”miRNA*” gives the sequence of any potential miRNA* sequences present in the predicted precursor.

Table A.3: miRNAs predicted in the GSM118373 454 leaf sRNA dataset [Kasschau et al., 2007] by `miRCat` using default settings

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
165a	1	78932	78952	-	885	TCGGACCAGGCTTCATCCCCC	21	2	144	40.97	-54	-37.5	0.01	GGAATGTTGTCTGGATCGAGG GAATGTTGTCTGGATCGAGGA

Continued on Next Page...

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
171b	1	3961367	3961387	-	1468	TTGAGCCGTGCCAATATCACG	21	2	117	45.3	-46.7	-39.91	0.01	CGAGATATTAGTGCGGTTCAA AGATATTAGTGCGGTTCAATC
472a	1	4182271	4182291	-	21	TGTATGTATGGTCGAAGTAGG	21	1	222	40.99	-91.6	-41.26	0.01	CCTACTCCGCCCATACCATAC
159b	1	6220806	6220826	+	2910	TTTGGATTGAAGGGAGCTCTT	21	1	188	39.89	-83.9	-44.63	0.01	GAGCTCCTTGAAGTTCAATGG
169h	1	6695535	6695555	-	1032	TAGCCAAGGATGACTTGCCTG	21	7	190	39.47	-80.5	-42.37	0.01	NO
864a	1	6740559	6740580	+	8	TTAAAGTCAATAATACCTTGAA	22	1	120	30.83	-39.17	-32.64	0.01	NO
394a	1	7058197	7058216	+	104	TTGGCATTCTGTCCACCTCC	20	2	198	36.36	-82.4	-41.62	0.01	AGGTGGGTATACTGCCAATA
395a	1	9363193	9363213	-	121	CTGAAGTGTTTGGGGGAACTC	21	3	103	44.66	-51.4	-49.9	0.01	NO
395b	1	9364527	9364547	+	62	CTGAAGTGTTTGGGGGGACTC	21	3	102	41.18	-45	-44.12	0.01	NO
395c	1	9367136	9367156	+	62	CTGAAGTGTTTGGGGGGACTC	21	3	111	42.34	-48.4	-43.6	0.01	NO
399a	1	10227151	10227171	+	89	TGCCAAAGGAGATTTGCCCTG	21	1	133	39.1	-57	-42.86	0.01	NO
400a	1	11785927	11785947	+	148	TATGAGAGTATTATAAGTCAC	21	2	155	32.9	-49.8	-32.13	0.01	AAGTGACTTATGATAATCTCA AGTGACTTATGATAATCTCAT
773a	1	13067291	13067312	+	29	TTTGCTTCCAGCTTTTGTCTCC	22	1	144	41.67	-61.2	-42.5	0.01	NO
161a	1	17829398	17829418	+	41003	TTGAAAGTGACTACATCGGGG	21	1	122	39.34	-53.5	-43.85	0.01	AACCCTGGTTTAGTCACTTTC ACCCTGGTTTAGTCACTTTCA
169d	1	20043224	20043244	-	622	TGAGCCAAGGATGACTTGCCG	21	4	143	37.76	-61.2	-42.8	0.01	CGGCAAGTTGACCTTGGCTCT GGCAAGTTGACCTTGGCTCTG
169e	1	20045254	20045274	+	622	TGAGCCAAGGATGACTTGCCG	21	4	152	39.47	-58.3	-38.36	0.01	GCAAGTTGACTTTGGCTCTGT
158b	1	20775940	20775958	+	18	CCCAAATGTAGACAAAGCA	19	2	95	40	-42	-44.21	0.01	NO
774a	1	22153670	22153690	+	23	TTGGTTACCCATATGGCCATC	21	1	139	30.94	-49.9	-35.9	0.01	TCAGATGGCTGTTTGGGTAAC TGGCTGTTTGGGTAACTAATA
Candidate 9	1	22376144	22376164	+	6	GTCATGGGGTATGATCGAATG	21	1	117	32.48	-51.1	-43.68	0.01	TTTGATCATTCTCCATGATAG
171c	1	22933760	22933780	-	1468	TTGAGCCGTGCCAATATCACG	21	2	95	37.89	-48.4	-50.95	0.01	AGATATTGGTGCGGTTCAATC
399b	1	23349052	23349072	-	537	TGCCAAAGGAGAGTTGCCCTG	21	2	162	38.27	-61.4	-37.9	0.01	GGGCGCCTCTCCATTGGCAGG
Candidate 7	1	23358770	23358790	-	7	TAGTGGAAGCAGCAACGAGAA	21	1	163	40.49	-73.8	-45.28	0.01	NO
Candidate 10	1	24557674	24557694	+	6	TTGTACAAATTTAAGTGTACG	21	1	135	27.41	-63.2	-46.81	0.01	TACACTTAGTTTTGTACAACA ACACTTAGTTTTGTACAACAT
157a	1	24916939	24916959	-	889	TTGACAGAAGATAGAGAGCAC	21	3	154	38.31	-61.2	-39.74	0.01	GCTCTCTAGCCTTCTGTCATC
157b	1	24924767	24924787	+	889	TTGACAGAAGATAGAGAGCAC	21	3	132	44.7	-57.06	-43.23	0.01	GCTCTCTAGCCTTCTGTCATC
839a	1	25282801	25282821	-	439	TACCAACCTTTCATCGTTCCC	21	1	235	42.55	-151.4	-64.43	0.01	GGAACGCATGAGAGGTTGGTA
395d	1	26273652	26273672	-	121	CTGAAGTGTTTGGGGGAACTC	21	3	115	40.87	-51.3	-44.61	0.01	NO
395e	1	26276449	26276469	-	121	CTGAAGTGTTTGGGGGAACTC	21	3	97	41.24	-41.5	-42.78	0.01	NO
395f	1	26277602	26277622	+	62	CTGAAGTGTTTGGGGGGACTC	21	3	123	42.28	-58.7	-47.72	0.01	NO
777a	1	26641746	26641767	+	14	TACGCATTGAGTTTCGTTGCTT	22	1	121	33.06	-42.6	-35.21	0.01	NO
159a	1	27716895	27716915	-	34285	TTTGGATTGAAGGGAGCTCTA	21	1	215	39.53	-80.9	-37.63	0.01	GAGCTCCTTAAAGTTCAAACA
Candidate 12	1	28060612	28060632	-	5	TTAACAATTTCAAGCAAAGAA	21	1	211	35.07	-87.6	-41.52	0.01	NO
394b	1	28573715	28573734	+	104	TTGGCATTCTGTCCACCTCC	20	2	238	37.39	-92.6	-38.91	0.01	AGGTGGGCATACTGCCAATA
402a	1	29021480	29021501	+	11	TTCGAGGCCTATTAAACCTCTG	22	1	205	34.63	-71.3	-34.78	0.01	NO
833a	1	29530158	29530179	+	7	TAGACCGATGTCAACAAACAAG	22	1	157	38.22	-87.3	-55.61	0.01	TTGTTTGTTGTACTCGGTCTAG TGTTTGTTGTACTCGGTCTAGT
840a	2	771384	771404	-	57	TTGTTTAGGTCCCTTAGTTTC	21	1	267	43.45	-125.19	-46.89	0.01	ACACTGAAGGACCTAAACTAA CACTGAAGGACCTAAACTAAC
398a	2	1040947	1040967	+	121	GGAGTGGCATGTGAACACATA	21	1	148	39.19	-56.74	-38.34	0.01	TTTGTGTTCTCAGGTCACCCC TTGTGTTCTCAGGTCACCCCT
Candidate 14	2	3537947	3537968	-	8	CTTGGTCACCAAGTTGGCTCGC	22	1	282	50.35	-101.8	-36.1	0.02	NO
156g	2	8419671	8419690	-	43	CGACAGAAGAGAGTGAGCAC	20	1	135	49.63	-61.4	-45.48	0.01	NO
779a	2	9567946	9567966	+	21	TGATTGGAAATTTCGTTGACT	21	1	163	32.52	-69.5	-42.64	0.01	NO
844a	2	9949282	9949302	-	10	TTATAAGCCATCTTACTAGTT	21	1	240	39.17	-86.3	-35.96	0.01	AATGGTAAGATTGCTTATAAG ATGGTAAGATTGCTTATAAGC
831a	2	10254470	10254491	+	6	TGATCTCTTCGTACTCTTCTTG	22	1	190	38.95	-88.4	-46.53	0.01	AGAAGCGTACAAGGAGATGAGG
156a	2	10683613	10683632	-	6713	TGACAGAAGAGAGTGAGCAC	20	6	103	46.6	-50	-48.54	0.01	TGCTCACTGCTCTTTCTGTC GCTCACTGCTCTTTCTGTCA
862a	2	10725185	10725205	+	8	TCCAATAGGTCGAGCATGTGC	21	1	319	37.62	-148.4	-46.52	0.01	NO
825a	2	11166852	11166872	+	365	TTCTCAAGAAGGTGCATGAAC	21	1	102	38.24	-42.1	-41.27	0.01	NO
172a	2	11949995	11950015	-	30497	AGAATCTTGATGATGCTGCAT	21	2	143	46.15	-64.3	-44.97	0.01	TGTGGCATCATCAAGATTCAC
Candidate 6	2	12277272	12277293	-	8	TAGAGGAAAATATAGAGTTGGG	22	2	193	31.61	-77.31	-40.06	0.01	ATGCGCAACTCTATATTTCCTC
399d	2	14450067	14450087	-	10	TGCCAAAGGAGATTTGCCCCG	21	1	167	40.12	-64.9	-38.86	0.01	NO
399f	2	14452163	14452183	+	38	TGCCAAAGGAGATTTGCCCGG	21	1	146	40.41	-56.8	-38.9	0.01	NO
Candidate 2	2	15618961	15618981	+	32	TGAGATGAAATCTTTGATTGG	21	1	131	37.4	-51.7	-39.47	0.01	NO
390a	2	16069049	16069069	+	723	AAGCTCAGGAGGGATAGCGCC	21	2	107	42.06	-51.6	-48.22	0.01	TGGCGCTATCCATCCTGAGTT GCGCTATCCATCCTGAGTTTC
160a	2	16347360	16347380	+	7028	TGCCTGGCTCCCTGTATGCCA	21	3	116	44.83	-54.9	-47.33	0.01	GCGTATGAGGAGCCATGCATA CGTATGAGGAGCCATGCATAT
393a	2	16659189	16659210	+	115	TCCAAAGGGATCGCATTGATCC	22	2	131	35.11	-46.6	-35.57	0.01	NO
319c	2	17036948	17036968	+	42	TTGGACTGAAGGGAGCTCCTT	21	1	195	40.51	-84	-43.08	0.01	GAAGGAGATTCTTTCAGTCCA GGAGATTCTTTCAGTCCAGTC
159c	2	19001899	19001916	+	298	TTTGGATTGAAGGGAGCT	18	3	197	43.15	-78.4	-39.8	0.01	NO
403a	2	19422147	19422168	+	1766	TGTTTTGTGCTTGAATCTAATT	22	1	128	32.81	-42.29	-33.04	0.01	ATTAGATTCACGCACAAACTCG
164a	2	19527840	19527860	+	11422	TGGAGAAGCAGGGCACGTGCA	21	2	113	46.9	-48.2	-42.65	0.01	CACGTACTTAACTTCTCCAAC
167c	3	1306756	1306776	-	150	TAAGCTGCCAGCATGATCTTG	21	1	159	34.59	-66.9	-42.08	0.01	TAGGTCATGCTGGTAGTTTCA AGGTCATGCTGGTAGTTTCAC
158a	3	3366354	3366373	-	2977	TCCCAAATGTAGACAAAGCA	20	1	125	40.8	-44.52	-35.62	0.01	CTTTGTCTACAATTTTGGAA
172c	3	3599797	3599817	-	8871	AGAATCTTGATGATGCTGCAG	21	2	174	37.36	-62.16	-35.72	0.01	TGTTGGAGCATCATCAAGATT GGAGCATCATCAAGATTCACA

Continued on Next Page. . .

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
823a	3	4496833	4496853	-	800	TGGGTGGTGATCATATAAGAT	21	1	97	32.99	-42.3	-43.61	0.01	NO
169f	3	4805806	4805826	-	622	TGAGCCAAGGATGACTTGCCG	21	4	184	42.39	-79.2	-43.04	0.01	GGCAAGTTGACCTTGGCTCTG GCAAGTTGACCTTGGCTCTGC
868a	3	6488405	6488425	+	5	CTTCTTAAGTGCTGATAATGC	21	1	193	35.23	-83.7	-43.37	0.01	NO
167a	3	8108097	8108117	+	52442	TGAAGCTGCCAGCATGATCTA	21	2	136	45.59	-57.1	-41.99	0.01	TAGATCATGTTCGCAGTTTCA GATCATGTTCGCAGTTTCACC
173a	3	8236168	8236189	+	164	TTCGCTTGCAGAGAGAAATCAC	22	1	123	39.84	-48.37	-39.33	0.01	TGATTCTCTGTGTAAGCGAAAG
Candidate 13	3	9871507	9871527	+	15	GAAGTCACACTCAGTGGATGC	21	1	122	50.82	-44.5	-36.48	0.06	NO
169i	3	9873343	9873363	-	1032	TAGCCAAGGATGACTTGCCTG	21	7	206	35.92	-88.5	-42.96	0.01	CAGGCAGTCTCCTTGGCTATC AGGCAGTCTCCTTGGCTATCC
169k	3	9876912	9876932	-	1032	TAGCCAAGGATGACTTGCCTG	21	7	240	35	-90	-37.5	0.01	CAGGCAGTCTCCTTGGCTATC AGGCAGTCTCCTTGGCTATCC
169l	3	9877277	9877297	-	1032	TAGCCAAGGATGACTTGCCTG	21	7	235	35.74	-92.24	-39.25	0.01	AGGCAGTCTCTTTGGCTATCT
169m	3	9879555	9879575	-	1032	TAGCCAAGGATGACTTGCCTG	21	7	208	37.5	-91.81	-44.14	0.01	CAGGCAGTCTCCTTGGCTATC AGGCAGTCTCCTTGGCTATCC
Candidate 4	3	11750277	11750297	-	12	TTAGTTGACGGAATTGTGGCG	21	1	117	40.17	-59.92	-51.21	0.01	NO
Candidate 8	3	17189474	17189494	+	7	TTGTGCGGTTCAAATAGTAAC	21	1	101	30.69	-37.9	-37.52	0.01	TTACTATTTGAATCGTACTGC
843a	3	17753132	17753152	+	30	TTTAGGTCGAGCTTCATTGGA	21	1	183	36.61	-87.1	-47.6	0.01	CTGTGAAGCTCGATCTAAAAG
171a	3	19084500	19084520	+	11649	TGATTGAGCCGCGCCAATATC	21	4	123	43.09	-46.6	-37.89	0.01	TATTGGCCTGGTTCACTCAGA
771a	3	19670359	19670380	-	153	TGAGCCTCTGTGGTAGCCCTCA	22	1	142	43.66	-57.3	-40.35	0.01	NO
172d	3	20598970	20598990	+	8871	AGAATCTTGATGATGCTGCAG	21	2	143	35.66	-51.2	-35.8	0.01	TATTGCAACATCTTCAAGATT GCAACATCTTCAAGATTCAGA
393b	3	20702636	20702657	+	115	TCCAAAGGGATCGCATTGATCC	22	2	174	35.06	-67.6	-38.85	0.01	NO
827a	3	22133788	22133808	-	29	TTAGATGACCATCAACAAACT	21	1	106	37.74	-37	-34.91	0.01	TTTGTTGATTGATATCTACAC
171b	3	22422519	22422539	+	11649	TGATTGAGCCGCGCCAATATC	21	4	139	53.96	-44.5	-32.01	0.24	NO
166b	3	22933276	22933296	+	5086	TCGGACCAGGCTTCATTCCCC	21	7	147	42.86	-62.16	-42.29	0.01	GGACTGTTGTCTGGCTCGAGG GACTGTTGTCTGGCTCGAGGA
167b	3	23417152	23417172	+	52442	TGAAGCTGCCAGCATGATCTA	21	2	142	44.37	-55.3	-38.94	0.01	NO
165b	4	369856	369876	-	885	TCGGACCAGGCTTCATCCCCC	21	2	182	37.91	-65.3	-35.88	0.01	GGAATGTTGTTTGGATCGAGG
826a	4	1340528	1340548	+	42	TAGTCCGGTTTTGGATACGTG	21	1	94	32.98	-40.6	-43.19	0.01	CGTGTCCAAAACCAGATTATC
447c	4	1523445	1523463	-	5	CCCCTTACAATGTCGAGTA	19	3	198	34.85	-72	-36.36	0.01	NO
447a	4	1528188	1528208	-	35	TGGGGACGAGATGTTTTGTTG	21	2	237	36.29	-111.2	-46.92	0.01	ACGAAGCATCTGTCCCCTGGT
447b	4	1535480	1535500	-	35	TGGGGACGAGATGTTTTGTTG	21	2	236	37.29	-105.8	-44.83	0.01	ACGAAGCATCTGTCCCCTGGT
397a	4	2625958	2625978	+	128	TCATTGAGTGCAGCGTTGATG	21	1	128	32.81	-51.4	-40.16	0.01	NO
850a	4	7845752	7845773	+	24	TAAGATCCGGACTACAACAAAG	22	1	255	33.73	-80.6	-31.61	0.01	NO
863a	4	7846826	7846846	+	11	TGCGATTGAGAGCAACAAGAC	21	1	199	23.12	-81.4	-40.9	0.01	NO
857a	4	7878194	7878214	-	23	TTTTGTATGTTGAAGGTGTAT	21	1	165	28.48	-37	-22.42	0.01	NO
397b	4	7878726	7878746	-	380	TCATTGAGTGCATCGTTGATG	21	1	126	28.57	-40.1	-31.83	0.01	NO
780a	4	8504140	8504160	-	22	TTCTTCGTGAATATCTGGCAT	21	1	192	36.46	-71.6	-37.29	0.01	NO
160b	4	9888999	9889019	+	7028	TGCCTGGCTCCCTGTATGCCA	21	3	89	46.07	-42.3	-47.53	0.01	NO
168a	4	10578663	10578683	+	5757	TCGCTTGGTGCAGGTCGGGAA	21	2	155	50.97	-71.4	-46.06	0.01	TGGATCCCGCCTTGCATCAAC GATCCCGCCTTGCATCAACTG
Candidate 1	4	11233793	11233813	-	40	TAGTAACAGAATTTGGTGTTA	21	1	122	31.97	-56.1	-45.98	0.01	NO
867a	4	11375398	11375418	+	39	TTGAACATGGTTTATTAGGAA	21	1	118	26.27	-39.1	-33.14	0.01	NO
169g	4	11483106	11483126	-	622	TGAGCCAAGGATGACTTGCCG	21	4	161	42.86	-72.3	-44.91	0.01	CGGCAAGTTGACCTTGGCTCT GGCAAGTTGACCTTGGCTCTG
Candidate 11	4	11962966	11962986	-	5	CCAAATTAATAGCAAAGTTTG	21	1	147	27.21	-49.6	-33.74	0.01	NO
845b	4	12214096	12214117	-	23	TCGCTCTGATACCAAATTGATG	22	1	184	37.5	-52.1	-28.32	0.01	NO
845a	4	12217467	12217487	-	448	CGGCTCTGATACCAATTGATG	21	1	164	45.12	-60.1	-36.65	0.01	NO
319a	4	12353119	12353139	+	420	TTGGACTGAAGGGAGCTCCCT	21	2	203	39.9	-83.09	-40.93	0.01	AGAGCTTCCTTGAGTCCATTC AGCTTCCTTGAGTCCATTCAC
828a	4	13847026	13847047	+	15	TCTTGCTTAAATGAGTATTCCA	22	1	127	38.58	-61.14	-48.14	0.01	AGATGCTCATTTGAGCAAGCAA
156b	4	15074951	15074970	+	6713	TGACAGAAGAGAGTGAGCAC	20	6	166	48.19	-88.9	-53.55	0.01	TGCTCACCTCTCTTTCTGTC
156c	4	15415497	15415516	-	6713	TGACAGAAGAGAGTGAGCAC	20	6	93	46.24	-46.7	-50.22	0.01	TGCTCACTGCTCTATCTGTC
164b	5	287583	287603	+	11422	TGGAGAAGCAGGGCACGTGCA	21	2	153	41.83	-65.5	-42.81	0.01	TCATGTGCCCATCTTCACCAT CATGTGCCCATCTTCACCATC
172b	5	1188212	1188232	-	30497	AGAATCTTGATGATGCTGCAT	21	2	117	38.46	-48.9	-41.79	0.01	GCAGCACCATTAAGATTCACA
162a	5	2634937	2634957	-	1026	TCGATAAACCTCTGCATCCAG	21	2	125	44	-53.3	-42.64	0.01	TGGAGGCAGCGGTTCATCGAT GGAGGCAGCGGTTCATCGATC
166d	5	2840709	2840729	+	5086	TCGGACCAGGCTTCATTCCCC	21	7	113	44.25	-41.6	-36.81	0.01	GGAATATTGTCTGGCTCGAGG
156d	5	3456714	3456733	-	6713	TGACAGAAGAGAGTGAGCAC	20	6	117	39.32	-57.4	-49.06	0.01	GCTCACTCTCTTTTTGTCAT
156e	5	3867214	3867233	+	6713	TGACAGAAGAGAGTGAGCAC	20	6	138	44.93	-68	-49.28	0.01	NO
848a	5	4479450	4479471	-	55	TGACATGGGACTGCCTAAGCTA	22	1	158	37.97	-62.5	-39.56	0.01	NO
398b	5	4691110	4691130	+	5007	TGTGTTCTCAGGTCACCCCTG	21	2	129	48.06	-58	-44.96	0.01	AGGGTTGATATGAGAACACAC GGGTTGATATGAGAACACACG
398c	5	4694781	4694801	+	5007	TGTGTTCTCAGGTCACCCCTG	21	2	157	44.59	-57.7	-36.75	0.01	AGGGTTGATATGAGAACACAC GGGTTGATATGAGAACACACG
162b	5	7740613	7740633	-	1026	TCGATAAACCTCTGCATCCAG	21	2	125	42.4	-53.8	-43.04	0.01	TGGAGGCAGCGGTTCATCGAT GGAGGCAGCGGTTCATCGATC
169b	5	8527512	8527532	+	237	TGCAGCCAAGGATGACTTGCC	21	2	121	38.84	-54	-44.63	0.01	GGCAAGTTGTCCTTCGGCTAC
860a	5	9098879	9098899	+	7	TCAATAGATTGGACTATGTAT	21	1	156	31.41	-70.9	-45.45	0.01	ATATATAGTCCAATCTATTGA
156f	5	9136129	9136148	+	6713	TGACAGAAGAGAGTGAGCAC	20	6	132	50	-64.12	-48.58	0.01	GCTCACTCTCTATCCGTCAC
164c	5	9852688	9852708	+	520	TGGAGAAGCAGGGCACGTGCG	21	1	116	48.28	-51.8	-44.66	0.01	CACGTGTTCTACTACTCCAAC
319b	5	16677697	16677717	-	420	TTGGACTGAAGGGAGCTCCCT	21	2	253	35.97	-104.2	-41.19	0.01	AGAGCTTTCTTCGGTCCACTC GAGCTTTCTTCGGTCCACTCA

Continued on Next Page...

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
166e	5	16792752	16792772	-	5086	TCGGACCAGGCTTCATTCCCC	21	7	143	37.06	-49.3	-34.48	0.01	GGGGAATGTTGTCTGGCACGA GGGAATGTTGTCTGGCACGAG
166f	5	17533605	17533625	+	5086	TCGGACCAGGCTTCATTCCCC	21	7	105	43.81	-42.1	-40.1	0.01	TGAATGATGCCTGGCTCGAGA
168b	5	18376100	18376120	-	5757	TCGCTTGGTGCAGGTCGGGAA	21	2	156	50	-68	-43.59	0.01	CCCGTCTTGTATCAACTGAAT
160c	5	19026385	19026405	-	7028	TGCCTGGCTCCCTGTATGCCA	21	3	117	47.01	-55	-47.01	0.01	GCGTACAAGGAGTCAAGCATG CGTACAAGGAGTCAAGCATGA
870a	5	21412818	21412838	-	5	TTAGAATGTGATGCAAAACTT	21	1	85	34.12	-42.2	-49.65	0.01	NO
Candidate 5	5	22214403	22214423	+	9	GCTTCTTGGAGATGTGACGAT	21	1	96	47.92	-32.3	-33.65	0.02	NO
390b	5	23654187	23654207	+	723	AAGCTCAGGAGGGATAGCGCC	21	2	134	43.28	-60.5	-45.15	0.01	TGGCGCTATCCATCCTGAGTT GCGCTATCCATCCTGAGTTCC
172e	5	24005792	24005812	+	834	GGAATCTTGATGATGCTGCAT	21	1	136	47.79	-62.5	-45.96	0.01	TGCAGCACCATTAAGATTCAC GCAGCACCATTAAGATTCACA
391a	5	24310386	24310406	+	549	TTCGCAGGAGAGATAGCGCCA	21	1	124	38.71	-43.6	-35.16	0.01	ACGGTATCTCTCCTACGTAGC
Candidate 3	5	24326846	24326866	-	20	TGCAAATCCAGTTCTTGTGTC	21	1	98	46.94	-29.44	-30.04	0.09	NO
399c	5	24979794	24979814	+	537	TGCCAAAGGAGAGTTGCCCTG	21	2	146	45.21	-67.69	-46.36	0.01	AGGGCATCTTTCTATTGGCAG GGGCATCTTTCTATTGGCAGG
166g	5	25522108	25522128	+	5086	TCGGACCAGGCTTCATTCCCC	21	7	122	43.44	-47.9	-39.26	0.01	GGAATGTTGTTTGGCTCGAGG

A.1.4 Calabrese *et al*. mouse embyonic stem cell Solexa results

Results from `miRCat` (using default parameters) when run on the Solexa *Mus musculus* embyonic stem cell sRNA set from Calabrese *et al*. [Calabrese et al., 2007].

Column ”miR” shows the miRBase accession of the miRNA (if availalbe), column ”chr” shows the *Mus musculus* REFSEQ chromosome the sequence maps to, column ”start” shows the start position of the predicted miRNA, column ”end” shows the start position of the predicted miRNA, column ”ori” shows the orientation of the miRNA (either Watson "+" or Crick "-" strand). Column ”abun” shows the abundance of the sRNA in the Solexa dataset, column ”seq” shows the sRNA sequence, column ”len” shows the length of the sRNA sequence, column ”g. hits” shows the number of times this sequence maps to the reference genome, column ”h. len” shows the length of the predicted miRNA precursor structure. Column ”G/C%” shows the percentage G/C composition of the miRNA hairpin sequence, column ”MFE” shows the minimum free energy of the folded

miRNA precursor sequence, column ”AMFE” shows the MFE per 100nt (therefore normalising the MFE), column ”p-value” shows the `randfold` p-value for the predicted hairpin precursor (using 100 randomisations). Column ”miRNA*” gives the sequence of any potential miRNA* sequences present in the predicted precursor.

Table A.4: miRNAs predicted in the mouse embyonic stem cell Solexa sRNA cloning [Calabrese et al., 2007] by `miRCat` using default settings

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
miR96	NC_000072	30119506	30119528	-	137	TTTGGCACTAGCACATTTTTGCT	23	1	118	53.39	-53.8	-45.59	0.01	NO
miR93	NC_000071	138606802	138606824	-	221	CAAAGTGCTGTTCGTGCAGGTAG	23	1	86	56.98	-45.3	-52.67	0.01	NO
miR92b	NC_000069	89031048	89031069	-	21	TATTGCACTCGTCCCGGCCTCC	22	1	110	70.91	-61.42	-55.84	0.01	NO
miR9	NC_000073	86650165	86650187	+	20	TCTTTGGTTATCTAGCTGTATGA	23	3	188	62.77	-96.9	-51.54	0.01	NO
miR9	NC_000069	88019535	88019557	+	20	TCTTTGGTTATCTAGCTGTATGA	23	3	108	45.37	-45.7	-42.31	0.01	NO
miR9	NC_000079	83878426	83878448	+	20	TCTTTGGTTATCTAGCTGTATGA	23	3	125	44	-53.2	-42.56	0.01	NO
miR872	NC_000070	94331897	94331918	+	49	TGAACTATTGCAGTAGCCTCCT	22	1	122	45.08	-49.39	-40.48	0.01	AAGGTTACTTGTTAGTTCAGGA
miR7a	NC_000073	86033181	86033203	+	71	TGGAAGACTAGTGATTTTGTTGT	23	2	86	51.16	-39.4	-45.81	0.01	NO
miR7a	NC_000079	58494202	58494224	-	71	TGGAAGACTAGTGATTTTGTTGT	23	2	91	40.66	-43.7	-48.02	0.01	CAACAAATCACAGTCTGCCATAT
miR758	NC_000078	110951068	110951089	+	18	TTTGTGACCTGGTCCACTAACC	22	1	81	51.85	-32.3	-39.88	0.01	NO
miR708	NC_000073	103397960	103397982	+	15	AAGGAGCTTACAATCTAGCTGGG	23	1	139	53.24	-58.6	-42.16	0.01	NO
miR674	NC_000068	117010887	117010909	+	14	GCACTGAGATGGGAGTGGTGTAA	23	1	90	55.56	-56.5	-62.78	0.01	NO
miR673	NC_000078	110810214	110810235	+	25	CTCACAGCTCTGGTCCTTGGAG	22	1	145	63.45	-68.4	-47.17	0.1	NO
miR672	NC_000086	101311567	101311589	-	103	TGAGGTTGGTGTACTGTGTGTGA	23	1	129	37.98	-51.03	-39.56	0.01	NO
miR669f	NC_000068	10388912	10388934	+	131	CATATACATACACACACACGTAT	23	2	134	42.54	-50.91	-37.99	0.02	AGTTGTGTGTGCATGTGCATGTG
miR669d	NC_000068	10390011	10390032	+	61	ACTTGTGTGTGCATGTATATGT	22	2	167	39.52	-63.4	-37.96	0.01	TACATATACATACACACCCATA
miR669d	NC_000068	10393285	10393306	+	61	ACTTGTGTGTGCATGTATATGT	22	2	179	38.55	-60.3	-33.69	0.01	TACATATACATACACACCCATA
miR669c	NC_000068	10430946	10430969	+	75	ATAGTTGTGTGTGGATGTGTGTAT	24	1	150	41.33	-61.3	-40.87	0.01	NO
miR669b	NC_000068	10389479	10389500	+	30	ATATACATACACACAAACATAT	22	7	104	36.54	-42.4	-40.77	0.01	AGTTTTGTGTGCATGTGCATGT
miR669	NC_000068	10398535	10398557	+	1203	ACATAACATACACACACACGTAT	23	13	105	36.19	-37.5	-35.71	0.01	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10408331	10408353	+	1203	ACATAACATACACACACACGTAT	23	13	122	39.34	-45.2	-37.05	0.01	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10413244	10413266	+	1203	ACATAACATACACACACACGTAT	23	13	122	39.34	-45.2	-37.05	0.02	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10400993	10401015	+	1203	ACATAACATACACACACACGTAT	23	13	122	39.34	-45.2	-37.05	0.01	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10403439	10403461	+	1203	ACATAACATACACACACACGTAT	23	13	122	39.34	-45.2	-37.05	0.01	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10405874	10405896	+	1203	ACATAACATACACACACACGTAT	23	13	122	39.34	-45.2	-37.05	0.01	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10425469	10425491	+	1203	ACATAACATACACACACACGTAT	23	13	105	36.19	-37.5	-35.71	0.02	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10423017	10423039	+	1203	ACATAACATACACACACACGTAT	23	13	122	39.34	-43.7	-35.82	0.01	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10420589	10420611	+	1203	ACATAACATACACACACACGTAT	23	13	122	39.34	-45.2	-37.05	0.01	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10435982	10436004	+	1203	ACATAACATACACACACACGTAT	23	13	114	38.6	-43.6	-38.25	0.01	TAGTTGTGTGTGCATGTTTATGT
miR669	NC_000068	10433885	10433907	+	1203	ACATAACATACACACACACGTAT	23	13	105	36.19	-37.5	-35.71	0.02	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10415706	10415728	+	1203	ACATAACATACACACACACGTAT	23	13	122	39.34	-45.2	-37.05	0.01	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10431848	10431870	+	1203	ACATAACATACACACACACGTAT	23	13	115	38.26	-40.8	-35.48	0.01	TAGTTGTGTGTGCATGTTCATGT
miR669	NC_000068	10396087	10396110	+	113	AGTTGTGTGTGCATGTTCATGTCT	24	15	122	38.52	-45.6	-37.38	0.01	NO
miR669	NC_000068	10410757	10410780	+	113	AGTTGTGTGTGCATGTTCATGTCT	24	15	129	38.76	-47.9	-37.13	0.01	NO
miR669	NC_000068	10418138	10418161	+	113	AGTTGTGTGTGCATGTTCATGTCT	24	15	181	39.78	-62.2	-34.36	0.01	NO
miR542	NC_000086	50402591	50402613	-	12	TGTGACAGATTGATAACTGAAAG	23	1	85	51.76	-38.8	-45.65	0.01	NO
miR541	NC_000078	110980632	110980656	+	129	AAGGGATTCTGATGTTGGTCACACT	25	1	107	42.06	-39.4	-36.82	0.01	NO
miR540	NC_000078	110824295	110824317	+	12	CAAGGGTCACCCTCTGACTCTGT	23	1	106	64.15	-67.2	-63.4	0.01	NO
miR539	NC_000078	110966346	110966368	+	30	GGAGAAATTATCCTTGGTGTGTT	23	1	135	40.74	-48.4	-35.85	0.01	NO
miR532	NC_000086	6825582	6825603	-	26	CATGCCTTGAGTGTAGGACCGT	22	1	183	47.54	-64.06	-35.01	0.01	NO
miR500	NC_000086	6814821	6814842	-	12	AATGCACCTGGGCAAGGGTTCA	22	1	124	52.42	-59.73	-48.17	0.01	NO

Continued on Next Page. . .

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
miR497	NC_000077	70048232	70048253	+	67	CAGCAGCACACTGTGGTTTGTA	22	1	87	62.07	-43.5	-50	0.01	NO
miR496	NC_000078	110977375	110977396	+	11	TGAGTATTACATGGCCAATCTC	22	1	79	41.77	-35.3	-44.68	0.01	GGTTGCCCATGGTGTGTTCATT
miR495	NC_000078	110957005	110957026	+	564	AAACAAACATGGTGCACTTCTT	22	1	84	35.71	-29.4	-35	0.01	GAAGTTGCCCATGTTATTTTTC
miR494	NC_000078	110953577	110953599	+	922	TGAAACATACACGGGAAACCTCT	23	1	85	35.29	-36.7	-43.18	0.01	GAGAGGTTGTCCGTGTTGTCTTC
miR493	NC_000078	110818453	110818474	+	17	TTGTACATGGTAGGCTTTCATT	22	1	101	55.45	-50.99	-50.49	0.01	NO
miR485	NC_000078	110973156	110973178	+	214	AGTCATACACGGCTCTCCTCTCT	23	1	113	48.67	-48.5	-42.92	0.01	NO
miR476e	NC_000068	10427359	10427380	+	50	TAAGTGTGAGCATGTATATGTG	22	1	146	43.15	-60.1	-41.16	0.01	CATATACATACACACACCTATA
miR467c	NC_000068	10395568	10395589	+	59	TAAGTGCGTGCATGTATATGTG	22	1	151	43.05	-55.4	-36.69	0.01	NO
miR467b	NC_000068	10402884	10402905	+	155	TAAGTGCCTGCATGTATATGCG	22	12	140	46.43	-52.2	-37.29	0.01	NO
miR467	NC_000068	10417657	10417678	+	597	ATATACATACACACACCTACAC	22	12	151	45.7	-64.5	-42.72	0.01	GTAAGTGCCTGCATGTATATGC
miR467	NC_000068	10410276	10410297	+	597	ATATACATACACACACCTACAC	22	12	151	45.7	-64.5	-42.72	0.01	GTAAGTGCCTGCATGTATATGC
miR467	NC_000068	10407812	10407833	+	597	ATATACATACACACACCTACAC	22	12	151	45.7	-64.5	-42.72	0.01	GTAAGTGCCTGCATGTATATGC
miR467	NC_000068	10420070	10420091	+	597	ATATACATACACACACCTACAC	22	12	151	46.36	-64.5	-42.72	0.01	GTAAGTGCCTGCATGTATATGC
miR467	NC_000068	10415187	10415208	+	597	ATATACATACACACACCTACAC	22	12	151	45.7	-64.5	-42.72	0.01	GTAAGTGCCTGCATGTATATGC
miR467	NC_000068	10412725	10412746	+	597	ATATACATACACACACCTACAC	22	12	151	45.03	-61.8	-40.93	0.01	GTAAGTGCCTGCATGTATATGC
miR467	NC_000068	10398018	10398039	+	597	ATATACATACACACACCTACAC	22	12	150	46	-63.6	-42.4	0.01	GTAAGTGCCTGCATGTATATGC
miR467	NC_000068	10424950	10424971	+	597	ATATACATACACACACCTACAC	22	12	151	45.03	-61.8	-40.93	0.01	GTAAGTGCCTGCATGTATATGC
miR467	NC_000068	10422499	10422520	+	597	ATATACATACACACACCTACAC	22	12	151	46.36	-64.5	-42.72	0.01	GTAAGTGCCTGCATGTATATGC
miR467	NC_000068	10429303	10429324	+	597	ATATACATACACACACCTACAC	22	12	151	45.7	-60.6	-40.13	0.01	ATAAGTGCGCGCATGTATATGC
miR467	NC_000068	10405355	10405376	+	597	ATATACATACACACACCTACAC	22	12	151	45.03	-61.8	-40.93	0.01	GTAAGTGCCTGCATGTATATGC
miR467	NC_000068	10400475	10400496	+	597	ATATACATACACACACCTACAC	22	12	151	45.03	-60.8	-40.26	0.01	GTAAGTGCCTGCATGTATATGC
miR466f	NC_000079	71245986	71246007	+	39	ACGTGTGTGTGCATGTGCATGT	22	5	128	44.53	-44.7	-34.92	0.06	NO
miR466f	NC_000068	10393591	10393612	+	39	ACGTGTGTGTGCATGTGCATGT	22	5	107	41.12	-50.6	-47.29	0.09	ATACACACACACATACACACGC
miR466f	NC_000068	10388582	10388603	+	39	ACGTGTGTGTGCATGTGCATGT	22	5	139	38.85	-64.5	-46.4	0.01	CATACACATACACACACACATA
miR466b	NC_000068	10425242	10425263	+	198	ATACATACACGCACACATAAGA	22	15	127	39.37	-52.7	-41.5	0.01	TTGATGTGTGTGTACATGTACA
miR466	NC_000068	10398307	10398329	+	236	TATACATACACGCACACATAAGA	23	14	131	38.93	-53.6	-40.92	0.01	TTGATGTGTGTGTACATGTACAT
miR466	NC_000068	10395895	10395917	+	236	TATACATACACGCACACATAAGA	23	14	131	39.69	-55.5	-42.37	0.01	TTGATGTGTGTGTACATGTACAT
miR466	NC_000068	10400765	10400787	+	236	TATACATACACGCACACATAAGA	23	14	131	38.93	-53.6	-40.92	0.01	TTGATGTGTGTGTACATGTACAT
miR466	NC_000068	10403211	10403233	+	236	TATACATACACGCACACATAAGA	23	14	127	39.37	-55.2	-43.46	0.01	TGATGTGTGTGTGCATGTACATA
miR466	NC_000068	10405646	10405668	+	236	TATACATACACGCACACATAAGA	23	14	112	40.18	-47.2	-42.14	0.01	TGATGTGTGTGTGCATGTACATA
miR466	NC_000068	10415478	10415500	+	236	TATACATACACGCACACATAAGA	23	14	131	39.69	-57.3	-43.74	0.01	TTGATGTGTGTGTACATGTACAT
miR466	NC_000068	10417946	10417968	+	236	TATACATACACGCACACATAAGA	23	14	131	38.93	-50.8	-38.78	0.01	TTGATGTGTGTGTACATGTACAT
miR466	NC_000068	10420361	10420383	+	236	TATACATACACGCACACATAAGA	23	14	131	39.69	-53.22	-40.63	0.01	TTGATGTGTGTGTACATGTACAT
miR466	NC_000068	10408103	10408125	+	236	TATACATACACGCACACATAAGA	23	14	131	39.69	-57.3	-43.74	0.01	TTGATGTGTGTGTACATGTACAT
miR466	NC_000068	10410565	10410587	+	236	TATACATACACGCACACATAAGA	23	14	131	38.93	-50.8	-38.78	0.01	TTGATGTGTGTGTACATGTACAT
miR466	NC_000068	10413016	10413038	+	236	TATACATACACGCACACATAAGA	23	14	112	40.18	-47.2	-42.14	0.01	TGATGTGTGTGTGCATGTACATA
miR466	NC_000068	10422789	10422811	+	236	TATACATACACGCACACATAAGA	23	14	131	39.69	-57.3	-43.74	0.01	TTGATGTGTGTGTACATGTACAT
miR466	NC_000068	10427685	10427707	+	236	TATACATACACGCACACATAAGA	23	14	130	39.23	-49.2	-37.85	0.02	ATGTGTGTGTACATGTACATGTG
miR466	NC_000068	10429594	10429616	+	236	TATACATACACGCACACATAAGA	23	14	131	39.69	-50.2	-38.32	0.01	TTATGTGTGTGTACATGTACATA
miR466	NC_000068	10433644	10433665	+	91	TATACATACACGCACACATAGA	22	2	157	42.04	-74.8	-47.64	0.01	TGTGTGTGCGTACATGTACATG
miR466	NC_000069	50925190	50925211	+	91	TATACATACACGCACACATAGA	22	2	134	29.1	-49.35	-36.83	0.01	NO
miR466	NC_000068	10436528	10436551	+	64	TGTGTGCATGTGCTTGTGTGTATG	24	4	107	39.25	-42.42	-39.64	0.03	NO
miR466	NC_000069	120047126	120047149	-	64	TGTGTGCATGTGCTTGTGTGTATG	24	4	131	33.59	-43	-32.82	0.01	NO
miR450a	NC_000086	50401516	50401537	-	13	TTTTGCGATGTGTTCCTAATAT	22	2	120	41.67	-41.8	-34.83	0.01	NO
miR450a	NC_000086	50401383	50401404	-	13	TTTTGCGATGTGTTCCTAATAT	22	2	143	38.46	-50.04	-34.99	0.01	NO
miR434	NC_000078	110832775	110832796	+	639	TTTGAACCATCACTCGACTCCT	22	1	82	45.12	-37.1	-45.24	0.01	AGCTCGACTCATGGTTTGAACC
miR433	NC_000078	110829991	110830012	+	190	ATCATGATGGGCTCCTCGGTGT	22	1	86	51.16	-35.3	-41.05	0.01	TACGGTGAGCCTGTCATTATTC
miR425	NC_000075	108471120	108471142	+	53	AATGACACGATCACTCCCGTTGA	23	1	95	56.84	-38.5	-40.53	0.01	NO
miR423	NC_000077	76891624	76891646	-	16	TGAGGGGCAGAGAGCGAGACTTT	23	1	122	50	-62	-50.82	0.01	AGCTCGGTCTGAGGCCCCTCAGT
miR412	NC_000078	110981513	110981535	+	19	TGGTCGACCAGCTGGAAAGTAAT	23	1	74	54.05	-26.3	-35.54	0.06	NO
miR411	NC_000078	110948435	110948455	+	784	TATGTAACACGGTCCACTAAC	21	1	82	45.12	-30	-36.59	0.01	ATAGTAGACCGTATAGCGTAC
miR410	NC_000078	110981974	110981994	+	129	AATATAACACAGATGGCCTGT	21	1	79	40.51	-34.1	-43.16	0.01	AGGTTGTCTGTGATGAGTTCG
miR382	NC_000078	110972027	110972048	+	93	AATCATTCACGGACAACACTTT	22	1	82	41.46	-36.2	-44.15	0.01	GAAGTTGTTCGTGGTGGATTCG
miR381	NC_000078	110965080	110965101	+	12	TATACAAGGGCAAGCTCTCTGT	22	1	107	40.19	-48	-44.86	0.01	NO
miR380	NC_000078	110950052	110950072	+	14	TATGTAGTATGGTCCACATCT	21	1	82	42.68	-31.8	-38.78	0.01	NO
miR379	NC_000078	110947312	110947332	+	226	TATGTAACATGGTCCACTAAC	21	1	85	42.35	-34.4	-40.47	0.01	NO
miR378	NC_000084	61557491	61557512	-	10	ACTGGACTTGGAGTCAGAAGGC	22	1	264	59.47	-155	-58.71	0.01	CTCCTGACTCCAGGTCCTGTGT
miR377	NC_000078	110978763	110978784	+	49	ATCACACAAAGGCAACTTTTGT	22	1	127	47.24	-48.6	-38.27	0.01	AGAGGTTGCCCTTGGTGAATTC
miR376c	NC_000078	110960980	110961000	+	78	AACATAGAGGAAATTTCACGT	21	1	149	40.27	-49.14	-32.98	0.01	NO

Continued on Next Page. . .

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
miR376b	NC_000078	110961718	110961738	+	232	ATCATAGAGGAACATCCACTT	21	1	86	34.88	-36.2	-42.09	0.01	GTGGATATTCCTTCTATGGTT
miR376	NC_000078	110962034	110962054	+	182	ATCGTAGAGGAAAATCCACGT	21	1	103	32.04	-29.01	-28.17	0.01	NO
miR374	NC_000086	100768451	100768472	-	78	ATATAATACAACCTGCTAAGTG	22	1	68	42.65	-43.3	-63.68	0.01	NO
miR370	NC_000078	110856480	110856503	+	14	CAGGTCACGTCTCTGCAGTTACAC	24	1	110	58.18	-51.4	-46.73	0.01	NO
miR369	NC_000078	110981676	110981696	+	482	AATAATACATGGTTGATCTTT	21	1	81	40.74	-36.6	-45.19	0.01	AGATCGACCGTGTTATATTCG
miR362	NC_000086	6819145	6819168	-	16	AATCCTTGGAACCTAGGTGTGAAT	24	1	84	52.38	-40.8	-48.57	0.01	NO
miR361	NC_000086	110188476	110188497	-	85	TTATCAGAATCTCCAGGGGTAC	22	1	84	46.43	-35.8	-42.62	0.01	NO
miR350	NC_000067	178702471	178702494	-	117	TTCACAAAGCCCATACACTTTCAC	24	1	98	44.9	-34.72	-35.43	0.01	NO
miR34c	NC_000075	50911181	50911203	-	18	AGGCAGTGTAGTTAGCTGATTGC	23	1	87	39.08	-32.9	-37.82	0.01	NO
miR345	NC_000078	110075198	110075219	+	12	TGCTGACCCCTAGTCCAGTGCT	22	1	109	59.63	-53.36	-48.95	0.01	NO
miR341	NC_000078	110849770	110849790	+	44	TCGATCGGTCGGTCGGTCAGT	21	1	131	56.49	-56.54	-43.16	0.01	NO
miR340	NC_000077	49883222	49883243	+	28	TTATAAAGCAATGAGACTGATT	22	1	124	37.1	-36.1	-29.11	0.01	CCGTCTCAGTTACTTTATAGCC
miR34	NC_000070	149442582	149442603	+	224	TGGCAGTGTCTTAGCTGGTTGT	22	1	112	47.32	-51.9	-46.34	0.01	CAATCAGCAAGTATACTGCCCT
miR339	NC_000071	139845664	139845684	-	23	TCCCTGTCCTCCAGGAGCTCA	21	1	82	68.29	-42.6	-51.95	0.02	NO
miR337	NC_000078	110824028	110824049	+	134	CGGCGTCATGCAGGAGTTGATT	22	1	135	58.52	-76.1	-56.37	0.01	TTCAGCTCCTATATGATGCCTT
miR335	NC_000072	30691354	30691375	+	17	TTTTTCATTATTGCTCCTGACC	22	1	115	41.74	-52.2	-45.39	0.01	TCAAGAGCAATAACGAAAAATG
miR33	NC_000081	82028557	82028576	+	38	GTGCATTGTAGTTGCATTGC	20	1	95	54.74	-52	-54.74	0.01	NO
miR329	NC_000078	110951751	110951772	+	75	AACACACCCAGCTAACCTTTTT	22	1	82	41.46	-35.7	-43.54	0.01	AGAGGTTTTCTGGGTCTCTGTT
miR322	NC_000086	50407483	50407504	-	20	CAGCAGCAATTCATGTTTTGGA	22	1	163	57.06	-80.42	-49.34	0.01	NO
miR32	NC_000070	56908144	56908165	-	21	TATTGCACATTACTAAGTTGCA	22	1	96	42.71	-44.1	-45.94	0.01	NO
miR31	NC_000070	88556517	88556539	-	31	AGGCAAGATGCTGGCATAGCTGT	23	1	166	50	-70.5	-42.47	0.01	TGCTATGCCAACATATTGCCATC
miR30e	NC_000070	120445263	120445286	-	129	TGTAAACATCCTTGACTGGAAGCT	24	1	97	51.55	-54.5	-56.19	0.01	NO
miR30d	NC_000081	68172817	68172840	-	23	TGTAAACATCCCCGACTGGAAGCT	24	1	128	46.88	-42.5	-33.2	0.01	NO
miR30c	NC_000067	23298553	23298576	+	1873	TGTAAACATCCTACACTCTCAGCT	24	2	88	44.32	-31.84	-36.18	0.01	NO
miR30c	NC_000070	120442188	120442211	-	1873	TGTAAACATCCTACACTCTCAGCT	24	2	107	49.53	-39.2	-36.64	0.05	NO
miR30b	NC_000081	68169031	68169052	-	1365	TGTAAACATCCTACACTCAGCT	22	1	86	46.51	-40.7	-47.33	0.01	NO
miR30a	NC_000067	23279113	23279136	+	77	TGTAAACATCCTCGACTGGAAGCT	24	1	96	52.08	-45.5	-47.4	0.01	NO
miR301b	NC_000082	17124511	17124535	-	48	CAGTGCAATGGTATTGTCAAAGCAT	25	1	135	55.56	-51	-37.78	0.07	NO
miR301	NC_000077	86926556	86926580	+	102	CAGTGCAATAGTATTGTCAAAGCAT	25	1	97	49.48	-40.5	-41.75	0.01	NO
miR300	NC_000078	110962573	110962594	+	35	TATGCAAGGGCAAGCTCTCTTC	22	1	175	54.86	-77.3	-44.17	0.01	NO
miR29c	NC_000067	196863794	196863815	+	63	TAGCACCATTTGAAATCGGTTA	22	1	86	45.35	-34.7	-40.35	0.01	NO
miR29b	NC_000072	31013025	31013047	-	105	TAGCACCATTTGAAATCAGTGTT	23	2	74	36.49	-28.62	-38.68	0.01	NO
miR29b	NC_000067	196863285	196863307	+	105	TAGCACCATTTGAAATCAGTGTT	23	2	175	44	-66.25	-37.86	0.01	NO
miR299	NC_000078	110948854	110948875	+	95	TGGTTTACCGTCCCACATACAT	22	1	83	45.78	-48.7	-58.67	0.01	GTATGTGGGACGGTAAACCGCT
miR297	NC_000068	10433349	10433370	+	226	ATATACATACACACATACCCAT	22	5	111	31.53	-38	-34.23	0.01	NO
miR297	NC_000068	10430691	10430712	+	226	ATATACATACACACATACCCAT	22	5	98	32.65	-38.8	-39.59	0.01	ATGTATGTGTGCATGTACATGT
miR297	NC_000072	31012673	31012694	-	297	TAGCACCATCTGAAATCGGTTA	22	1	100	40	-36.94	-36.94	0.01	ACTGATTTCTTTTGGTGTTCAG
miR294	NC_000073	3220656	3220677	+	1336	ACTCAAAATGGAGGCCCTATCT	22	1	135	42.22	-43.1	-31.93	0.01	AAAGTGCTTCCCTTTTGTGTGT
miR293	NC_000073	3220391	3220412	+	4868	AGTGCCGCAGAGTTTGTAGTGT	22	1	91	41.76	-32	-35.16	0.01	TACTCAAACTGTGTGACATTTT
miR292	NC_000073	3219240	3219262	+	1684	AAGTGCCGCCAGGTTTTGAGTGT	23	1	122	47.54	-54.2	-44.43	0.01	NO
miR291a	NC_000073	3218933	3218954	+	1463	CATCAAAGTGGAGGCCCTCTCT	22	1	82	53.66	-44.3	-54.02	0.01	AAAGTGCTTCCACTTTGTGTGC
miR290	NC_000073	3218640	3218662	+	1507	ACTCAAACTATGGGGGCACTTTT	23	1	167	48.5	-61.8	-37.01	0.01	TTTAAAAAGTGCCGCCTAGTTTT
miR28	NC_000082	24827954	24827975	+	87	AAGGAGCTCACAGTCTATTGAG	22	1	98	51.02	-39.8	-40.61	0.01	CACTAGATTGTGAGCTGCTGGA
miR27b	NC_000079	63402068	63402088	+	329	TTCACAGTGGCTAAGTTCTGC	21	1	100	47	-50.8	-50.8	0.01	NO
miR26b	NC_000067	74440898	74440919	+	296	TTCAAGTAATTCAGGATAGGTT	22	1	122	60.66	-67.5	-55.33	0.01	NO
miR26a	NC_000075	118940929	118940950	+	686	TTCAAGTAATCCAGGATAGGCT	22	2	139	65.47	-78.3	-56.33	0.01	NO
miR26a	NC_000076	126432599	126432620	+	686	TTCAAGTAATCCAGGATAGGCT	22	2	113	54.87	-53.9	-47.7	0.01	NO
miR25	NC_000071	138606560	138606581	-	837	CATTGCACTTGTCTCGGTCTGA	22	1	91	63.74	-40	-43.96	0.07	NO
miR24	NC_000079	63402559	63402580	+	433	TGGCTCAGTTCAGCAGGAACAG	22	2	85	57.65	-29.92	-35.2	0.03	NO
miR222	NC_000086	18724029	18724049	-	21	AGCTACATCTGGCTACTGGGT	21	1	106	51.89	-53	-50	0.01	NO
miR221	NC_000086	18723433	18723455	-	195	AGCTACATTGTCTGCTGGGTTTC	23	1	122	46.72	-47.86	-39.23	0.01	NO
miR221	NC_000079	63401837	63401859	+	221	ATCACATTGCCAGGGATTACCAC	23	1	181	53.59	-66.44	-36.71	0.08	NO
miR22	NC_000077	75277274	75277295	+	70	AAGCTGCCAGTTGAAGAACTGT	22	1	119	57.98	-49	-41.18	0.01	AGTTCTTCAGTGGCAAGCTTTA
miR219	NC_000083	34161995	34162017	-	68	TGATTGTCCAAACGCAATTCTCG	23	1	129	68.99	-77.8	-60.31	0.01	NO
miR210	NC_000073	148407306	148407327	-	256	CTGTGCGTGTGACAGCGGCTGA	22	1	117	70.09	-58.4	-49.91	0.05	NO
miR21	NC_000077	86397621	86397643	-	3411	TAGCTTATCAGACTGATGTTGAC	23	1	101	45.54	-46.2	-45.74	0.01	NO
miR205	NC_000067	195333697	195333718	-	52	TCCTTCATTCCACCGGAGTCTG	22	1	134	53.73	-56.2	-41.94	0.02	NO
miR200c	NC_000072	124668342	124668364	-	60	TAATACTGCCGGGTAATGATGGA	23	1	92	58.7	-55.9	-60.76	0.01	NO
miR200b	NC_000070	155429793	155429815	-	420	TAATACTGCCTGGTAATGATGAC	23	1	82	56.1	-40.3	-49.15	0.01	CATCTTACTGGGCAGCATTGGAT

Continued on Next Page. . .

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
miR200a	NC_000070	155429019	155429041	-	362	TAACACTGTCTGGTAACGATGTT	23	1	88	51.14	-42.7	-48.52	0.01	CATCTTACCGGACAGTGCTGGAT
miR19a	NC_000080	115443270	115443292	+	1409	TGTGCAAATCTATGCAAAACTGA	23	1	95	41.05	-40.3	-42.42	0.01	TAGTTTTGCATAGTTGCACTACA
miR195	NC_000077	70048564	70048585	+	30	TAGCAGCACAGAAATATTGGCA	22	1	118	53.39	-55.8	-47.29	0.01	NO
miR194	NC_000067	187137204	187137225	+	32	TGTAACAGCAACTCCATGTGGA	22	2	82	53.66	-46.4	-56.59	0.01	NO
miR191	NC_000075	108470656	108470677	+	116	CAACGGAATCCCAAAAGCAGCT	22	1	101	56.44	-49.6	-49.11	0.01	GCTGCACTTGGATTTCGTTCCC
miR190	NC_000075	67084506	67084528	-	22	TGATATGTTTGATATATTAGGTT	23	1	149	40.27	-46.1	-30.94	0.01	NO
miR18b	NC_000086	50095558	50095580	-	57	TAAGGTGCATCTAGTGCTGTTAG	23	1	137	43.07	-36.5	-26.64	0.05	NO
miR185	NC_000082	18327531	18327552	-	19	TGGAGAGAAAGGCAGTTCCTGA	22	1	82	60.98	-53.1	-64.76	0.01	NO
miR183	NC_000072	30119711	30119732	-	213	TATGGCACTGGTAGAATTCACT	22	1	113	51.33	-42.7	-37.79	0.01	GTGAATTACCGAAGGGCCATAA
miR182	NC_000072	30115962	30115986	-	360	TTTGGCAATGGTAGAACTCACACCG	25	1	102	52.94	-40.94	-40.14	0.03	NO
miR181d	NC_000074	86702657	86702680	-	54	AACATTCATTGTTGTCGGTGGGTT	24	1	97	56.7	-42.7	-44.02	0.01	NO
miR181c	NC_000074	86702821	86702844	-	421	AACATTCAACCTGTCGGTGAGTTT	24	1	87	58.62	-44	-50.57	0.01	NO
miR181a	NC_000067	139863045	139863069	+	17	AACATTCAACGCTGTCGGTGAGTTT	25	2	154	46.75	-43.62	-28.32	0.01	NO
miR181a	NC_000068	38708261	38708285	+	17	AACATTCAACGCTGTCGGTGAGTTT	25	2	90	45.56	-35.33	-39.26	0.01	NO
miR15b	NC_000069	68813697	68813718	+	1504	TAGCAGCACATCATGGTTTACA	22	1	84	36.9	-28.5	-33.93	0.01	GCGAATCATTATTTGCTGCTCT
miR151	NC_000081	73085284	73085305	-	182	TCGAGGAGCTCACAGTCTAGTA	22	1	133	57.89	-69.2	-52.03	0.01	CTAGACTGAGGCTCCTTGAGGA
miR150	NC_000073	52377132	52377153	+	40	TCTCCCAACCCTTGTACCAGTG	22	1	112	58.93	-54.3	-48.48	0.02	NO
miR148a	NC_000072	51219828	51219849	-	129	TCAGTGCACTACAGAACTTTGT	22	1	82	43.9	-35.7	-43.54	0.01	AAAGTTCTGAGACACTCCGACT
miR140	NC_000074	110075149	110075170	+	22	CAGTGGTTTTACCCTATGGTAG	22	1	98	53.06	-50.9	-51.94	0.01	ACCACAGGGTAGAACCACGGAC
miR136	NC_000078	110833541	110833563	+	23	ACTCCATTTGTTTTGATGATGGA	23	1	144	48.61	-62.8	-43.61	0.01	NO
miR135b	NC_000067	134094680	134094702	+	62	TATGGCTTTTCATTCCTATGTGA	23	1	95	54.74	-51.9	-54.63	0.01	NO
miR134	NC_000078	110972355	110972376	+	25	TGTGACTGGTTGACCAGAGGGG	22	1	71	60.56	-34.5	-48.59	0.05	NO
miR130b	NC_000082	17124164	17124185	-	271	CAGTGCAATGATGAAAGGGCAT	22	1	82	56.1	-32.1	-39.15	0.03	CACTCTTTCCCTGTTGCACTAC
miR130	NC_000068	84581273	84581294	-	1003	CAGTGCAATGTTAAAAGGGCAT	22	1	137	56.93	-77	-56.2	0.01	GCTCTTTTCACATTGTGCTACT
miR128	NC_000067	130098981	130099001	+	187	TCACAGTGAACCGGTCTCTTT	21	2	80	50	-35.6	-44.5	0.01	NO
miR128	NC_000075	112021148	112021168	-	187	TCACAGTGAACCGGTCTCTTT	21	2	85	55.29	-40.5	-47.65	0.01	NO
miR127	NC_000078	110831098	110831119	+	321	TCGGATCCGTCTGAGCTTGGCT	22	1	86	54.65	-38.6	-44.88	0.01	NO
miR126	NC_000068	10435079	10435100	+	29	AGTTTTGTGTGCATGTGCATGT	22	3	98	31.63	-40.6	-41.43	0.01	CATATACATCCACACAAACATA
miR126	NC_000068	10434436	10434457	+	29	AGTTTTGTGTGCATGTGCATGT	22	3	98	35.71	-38.2	-38.98	0.03	CATATACATCCACACAAACATA
miR125b	NC_000082	77646524	77646545	+	21	TCCCTGAGACCCTAACTTGTGA	22	2	79	48.1	-41.6	-52.66	0.01	NO
miR125b	NC_000075	41390023	41390044	+	21	TCCCTGAGACCCTAACTTGTGA	22	2	99	54.55	-46.5	-46.97	0.01	NO
miR125a	NC_000083	17967781	17967804	+	26	TCCCTGAGACCCTTTAACCTGTGA	24	1	84	65.48	-44.5	-52.98	0.01	NO
miR124	NC_000068	180628788	180628809	+	75	TAAGGCACGCGGTGAATGCCAA	22	3	93	56.99	-40.3	-43.33	0.01	CGTGTTCACAGCGGACCTTGAT
miR124	NC_000069	17695723	17695744	+	75	TAAGGCACGCGGTGAATGCCAA	22	3	107	50.47	-44.6	-41.68	0.01	CGTGTTCACAGCGGACCTTGAT
miR124	NC_000080	65209546	65209567	+	75	TAAGGCACGCGGTGAATGCCAA	22	3	101	50.5	-42.42	-42	0.01	CGTGTTCACAGCGGACCTTGAT
miR1193	NC_000078	110953960	110953980	+	31	TAGGTCACCCGTTTTACTATC	21	1	108	52.78	-47.8	-44.26	0.01	ATGGTAGACCGGTGACGTACA
miR103	NC_000077	35595949	35595971	+	94	AGCAGCATTGTACAGGGCTATGA	23	2	84	47.62	-35.3	-42.02	0.02	NO
miR103	NC_000068	131113839	131113861	+	94	AGCAGCATTGTACAGGGCTATGA	23	2	135	51.85	-53.4	-39.56	0.01	AGCTTCTTTACAGTGCTGCCTTG
miR101b	NC_000085	29209828	29209849	+	46	GTACAGTACTGTGATAGCTGAA	22	1	160	48.75	-59.7	-37.31	0.01	NO
miR101a	NC_000070	101019562	101019583	-	31	GTACAGTACTGTGATAACTGAA	22	1	81	50.62	-45.4	-56.05	0.01	TCAGTTATCACAGTGCTGATGC
miR543	NC_000078	110955514	110955535	+	1382	AAACATTCGCGGTGCACTTCTT	22	1	131	48.09	-45.5	-34.73	0.01	GAAGTTGCCCGCGTGTTTTTCG
N/A	NC_000073	87297760	87297781	-	14	AGCCTTTAATTTCAGTACTTGG	22	1	253	50.99	-157.1	-62.09	0.01	NO
miR669	NC_000068	10390616	10390637	+	30	AGTTGTGTGTGCATGTATATGT	22	1	122	36.89	-40.1	-32.87	0.05	TACATATACATACACACCCATA
miR467	NC_000086	113878990	113879011	-	25	ATATACATACACACACCTATAT	22	3	113	32.74	-56.1	-49.65	0.01	NO
miR297c	NC_000067	171984192	171984213	+	171	ATGTATGTGTGCATGTACATGT	22	7	98	37.76	-27.2	-27.76	0.03	NO
miR297c	NC_000076	80436056	80436077	-	171	ATGTATGTGTGCATGTACATGT	22	7	180	36.67	-81.4	-45.22	0.06	NO
miR669	NC_000076	82774935	82774955	-	10	CATATACATACACACACACGT	21	4	175	23.43	-77.8	-44.46	0.01	NO
N/A	NC_000079	21988799	21988819	-	30	CTCACCTGGAGCATGTTTTCT	21	1	60	51.67	-22.3	-37.17	0.02	NO
miR466d	NC_000068	10435386	10435407	+	66	GTGTGTGCGTACATGTACATGT	22	2	136	46.32	-74.7	-54.93	0.01	TATACATGAGAGCATACATAGA
miR466a	NC_000078	89981901	89981921	-	21	TATGTGTGTGTACATGTACAT	21	13	104	39.42	-33.3	-32.02	0.01	NO
miR466a	NC_000075	22519606	22519628	-	23	TATGTGTGTGTACATGTACATAT	23	5	60	36.67	-29.5	-49.17	0.01	NO
miR466a	NC_000068	174237240	174237262	-	23	TATGTGTGTGTACATGTACATAT	23	5	138	43.48	-40.6	-29.42	0.05	NO
miR466a	NC_000070	77513268	77513290	+	23	TATGTGTGTGTACATGTACATAT	23	5	132	34.09	-49.39	-37.42	0.01	NO
miR126	NC_000068	26446922	26446943	+	29	TCGTACCGTGAGTAATAATGCG	22	1	77	48.05	-33.7	-43.77	0.01	NO

A.1.5 Dalmay group chicken sRNA Solexa/Illumina results

Results from `miRCat` (using default parameters) when run on the Solexa *Gallus gallus* sRNA set from the Dalmay group.

Column ”miR” shows the miRBase accession of the miRNA (if availalbe), column ”chr” shows the chromosome the sequence maps to, column ”start” shows the start position of the predicted miRNA, column ”end” shows the start position of the predicted miRNA, column ”ori” shows the orientation of the miRNA (either Watson "+" or Crick "-" strand). Column ”abun” shows the abundance of the sRNA in the Solexa dataset, column ”seq” shows the sRNA sequence, column ”len” shows the length of the sRNA sequence, column ”g. hits” shows the number of times this sequence maps to the reference genome, column ”h. len” shows the length of the predicted miRNA precursor structure. Column ”G/C%” shows the percentage G/C composition of the miRNA hairpin sequence, column ”MFE” shows the minimum free energy of the folded miRNA precursor sequence, column ”AMFE” shows the MFE per 100nt (therefore normalising the MFE), column ”p-value” shows the `randfold` p-value for the predicted hairpin precursor (using 100 randomisations). Column ”miRNA*” gives the sequence of any potential miRNA* sequences present in the predicted precursor.

Table A.5: miRNAs predicted in the Dalmay group chicken Solexa sRNA cloning by `miRCat` using default settings

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
miR-29	1	3235362	3235381	+	29	TAGCACCATTTGAAATCAGT	20	2	76	36.84	-30.02	-39.5	0.01	NO
N/A	1	15746124	15746141	-	10	AGCGGCGCGGTAGGAGCA	18	1	146	70.55	-80.9	-55.41	0.01	NO
miR-33	1	51372325	51372345	-	29	GTGCATTGTAGTTGCATTGCA	21	1	69	47.83	-37.6	-54.49	0.01	NO
N/A	1	62970293	62970314	+	20	GTTTGGCTGTAGGCATGTGGGT	22	1	114	51.75	-45.3	-39.74	0.01	NO
let-7	1	73421275	73421296	+	95	TGAGGTAGTAGGTTGTATAGTT	22	4	85	42.35	-38.3	-45.06	0.01	NO
miR-99	1	102424345	102424365	+	165	AACCCGTAGATCCGATCTTGT	21	1	124	44.35	-50.1	-40.4	0.01	NO
let-7	1	102425096	102425118	+	45	TGAGGTAGTAGGTTGTATGGTTT	23	1	118	50.85	-49.9	-42.29	0.01	NO
miR-125	1	102457663	102457684	+	408	TCCCTGAGACCCTAACTTGTGA	22	1	89	50.56	-37.6	-42.25	0.01	NO

Continued on Next Page...

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
N/A	1	104458572	104458592	-	78	CGAGAAGACGGTCGAACTTGA	21	1	111	60.36	-60.3	-54.32	0.01	NO
N/A	1	104460527	104460550	-	25	CGGGGATCGGGCGCGCCTCTCCGT	24	2	132	83.33	-94.5	-71.59	0.01	NO
miR-222	1	114216044	114216066	+	57	CGCTCAGTAGTCAGTGTAGATTC	23	2	75	45.33	-32.8	-43.73	0.01	NO
miR-222	1	114218439	114218461	+	57	CGCTCAGTAGTCAGTGTAGATTC	23	2	75	45.33	-32.8	-43.73	0.01	NO
N/A	1	130934912	130934932	+	25	TGCATTGCGACGGGTTATATC	21	1	82	46.34	-36.1	-44.02	0.01	NO
miR-92	1	152248079	152248100	-	882	TATTGCACTTGTCCCGGCCTGT	22	1	86	54.65	-43.1	-50.12	0.01	NO
miR-19	1	152248505	152248524	-	65	TGTGCAAATCTATGCAAAAC	20	1	82	41.46	-35.9	-43.78	0.01	NO
miR-17	1	152248830	152248852	-	185	CAAAGTGCTTACAGTGCAGGTAG	23	1	85	41.18	-33	-38.82	0.01	NO
N/A	1	170154883	170154903	+	5	CGGGAGGGGAGGGAGGGCGGG	21	1	104	75	-72.9	-70.1	0.01	NO
miR-16	1	173700401	173700421	-	74	TAGCAGCACGTAAATATTGGT	21	2	71	35.21	-26.39	-37.17	0.01	NO
miR-26	2	4467525	4467546	+	754	TTCAAGTAATCCAGGATAGGCT	22	1	145	51.72	-63.9	-44.07	0.01	NO
miR-489	2	23068888	23068908	-	164	TGACATCATATGTACGGCTGC	21	1	98	46.94	-43	-43.88	0.01	NO
miR-148	2	32053546	32053567	-	748	TCAGTGCACTACAGAACTTTGT	22	1	82	51.22	-36.1	-44.02	0.01	NO
miR-196	2	32586206	32586227	-	29	TAGGTAGTTTCATGTTGTTGGG	22	3	69	44.93	-27.6	-40	0.01	NO
miR-128	2	45549227	45549247	+	671	TCACAGTGAACCGGTCTCTTT	21	2	93	54.84	-43.8	-47.1	0.01	NO
miR-32	2	86506497	86506516	-	24	ATATTGCACATTACTAAGTT	20	1	98	44.9	-45.2	-46.12	0.01	NO
miR-133	2	105670371	105670392	-	45	TTTGGTCCCCTTCAACCAGCTG	22	3	86	43.02	-36.6	-42.56	0.01	NO
miR-1	2	105673494	105673515	-	694	TGGAATGTAAAGAAGTATGTAT	22	2	91	38.46	-35.94	-39.49	0.01	NO
N/A	2	106776399	106776416	-	13	TTCTGTAGACTGTTTGAC	18	2	112	37.5	-42.3	-37.77	0.01	NO
miR-124	2	118524208	118524225	+	181	TAAGGCACGCGGTGAATG	18	2	84	51.19	-35.6	-42.38	0.01	NO
miR-30	2	148331648	148331669	-	62	TGTAAACATCCTACACTCAGCT	22	1	74	44.59	-30.2	-40.81	0.01	NO
miR-30	2	148337299	148337321	-	263	TGTAAACATCCCCGACTGGAAGC	23	1	79	49.37	-35.1	-44.43	0.01	NO
N/A	3	4348337	4348358	+	315	ACGGGACAGTGCTGAAGACTAC	22	1	138	63.77	-58.9	-42.68	0.01	NO
N/A	3	23002020	23002038	+	14	TCCTGCAGAAGGTGCGGCT	19	1	81	67.9	-41.8	-51.6	0.01	NO
miR-456	3	32679732	32679751	-	379	CAGGCTGGTTAGATGGTTGT	20	1	79	49.37	-33.5	-42.41	0.01	NO
miR-30	3	85102286	85102307	+	14276	CTTTCAGTCGGATGTTTGCAGC	22	1	97	53.61	-46.4	-47.84	0.01	TGTAAACATCCTCGACTGGAAG
miR-30	3	85126899	85126920	+	158	CTGGGAGAAGGCTGTTTACTCT	22	1	123	47.97	-49.5	-40.24	0.01	NO
miR-133	3	110384948	110384969	-	41	TTTGGTCCCCTTCAACCAGCTA	22	1	79	55.7	-37.2	-47.09	0.01	NO
miR-206	3	110390449	110390470	-	259	TGGAATGTAAGGAAGTGTGTGG	22	1	123	45.53	-51.6	-41.95	0.01	NO
miR-233	4	233007	233027	+	51	TGTCAGTTTGTCAAATACCCC	21	1	99	50.51	-45.2	-45.66	0.01	NO
miR-20	4	3970100	3970121	-	123	CAAAGTGCTCATAGTGCAGGTA	22	1	79	45.57	-33.1	-41.9	0.01	NO
miR-15	4	4049101	4049120	-	211	TAGCAGCACATCATGGTTTG	20	2	82	48.78	-37.9	-46.22	0.01	NO
miR-302	4	58651323	58651343	+	33	TTTAACATGGAGGTGCTTTCT	21	1	82	39.02	-33.02	-40.27	0.01	NO
miR-302	4	58651617	58651636	+	28	AGTGCTTCCATGTTTCAGTG	20	1	80	48.75	-39.7	-49.63	0.01	NO
miR-302	4	58651884	58651902	+	31	ACTTAAATGTGGATGTGCT	19	1	84	39.29	-30.6	-36.43	0.01	NO
miR-107	4	91906901	91906921	-	833	AGCAGCATTGTACAGGGCTAT	21	3	84	46.43	-37.2	-44.29	0.01	NO
miR-146	4	92169345	92169366	-	1596	TGAGAACTGAATTCCATGGACT	22	2	143	44.76	-60.09	-42.02	0.01	NO
N/A	5	26613484	26613505	-	27	TCAGAAAAGGATATGAATTGTC	22	1	79	39.24	-24.2	-30.63	0.01	NO
N/A	5	33777672	33777692	-	32	ACTAAGGACAGAGGAACGGAG	21	1	103	39.81	-31.52	-30.6	0.01	NO
N/A	5	45344529	45344546	-	26	GTCGTCGGGATGGAGTTT	18	1	69	59.42	-32.9	-47.68	0.01	NO
N/A	5	60284030	60284049	-	26	CGGGGCGGCTGTGAGCTGAG	20	1	65	75.38	-41.3	-63.54	0.01	NO
miR-107	6	20487975	20487995	-	833	AGCAGCATTGTACAGGGCTAT	21	3	84	45.24	-34	-40.48	0.01	NO
miR-202	6	22813083	22813103	+	152	TTCCTATGCATATACTTCTTT	21	1	78	41.03	-35.9	-46.03	0.01	NO
miR-146	6	24570077	24570094	+	21	TTGAGAACTGAATTCCAT	18	2	83	46.99	-42	-50.6	0.01	NO
N/A	6	34416337	34416357	-	39	TTAAGAGTAGGGATTCTGTTC	21	1	74	41.89	-26.2	-35.41	0.01	NO
miR-10	7	17389110	17389131	-	113106	TACCCTGTAGAACCGAATTTGT	22	1	151	40.4	-60.5	-40.07	0.01	NO
miR-375	7	23901164	23901185	+	30	TTTGTTCGTTCGGCTCGCGTTA	22	1	82	60.98	-43.8	-53.41	0.01	NO
miR-128	7	32228199	32228219	+	671	TCACAGTGAACCGGTCTCTTT	21	2	80	46.25	-35.7	-44.63	0.01	NO
miR-181	8	2001580	2001597	+	43	AACATTCAACGCTGTCGG	18	2	154	46.1	-44.61	-28.97	0.01	NO
miR-181	8	2001764	2001787	+	294	AACATTCATTGCTGTCGGTGGGTT	24	2	160	45.63	-54.1	-33.81	0.01	NO
N/A	8	3838518	3838538	-	28	TGGGTCCCGGCATGCTGCACT	21	1	89	64.04	-56.3	-63.26	0.01	NO
miR-199	8	4732840	4732861	+	149	ACAGTAGTCTGCACATTGGTTA	22	2	81	51.85	-36.7	-45.31	0.01	NO
miR-101	8	29051925	29051946	-	233	GTACAGTACTGTGATAACTGAA	22	2	81	49.38	-43.1	-53.21	0.01	NO
miR-551	9	21966435	21966452	-	25	GCGACCCATACTTGGTTT	18	1	113	56.64	-57.1	-50.53	0.01	NO
miR-16	9	23742848	23742868	-	74	TAGCAGCACGTAAATATTGGT	21	2	85	41.18	-32.3	-38	0.01	NO
miR-15	9	23743019	23743038	-	211	TAGCAGCACATCATGGTTTG	20	2	118	38.98	-39.12	-33.15	0.01	NO
N/A	10	1811809	1811828	-	19	TGCAGTGACGTCTCTTCCCC	20	1	64	71.88	-39.7	-62.03	0.01	NO
N/A	10	11522320	11522338	+	16	CAGGCGAGGGCGGGAGGGC	19	1	81	76.54	-54.9	-67.78	0.01	NO
miR-184	10	22146297	22146318	+	1168	TGGACGGAGAACTGATAAGGGT	22	1	86	58.14	-44.7	-51.98	0.01	NO

Continued on Next Page...

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
miR-140	11	21030701	21030722	+	1487	ACCACAGGGTAGAACCACGGAC	22	1	78	53.85	-44.6	-57.18	0.01	NO
let-7	12	6302556	6302577	-	386	TGAGGTAGTAGATTGTATAGTT	22	1	146	41.1	-63.49	-43.49	0.01	NO
let-7	12	6302967	6302988	-	95	TGAGGTAGTAGGTTGTATAGTT	22	4	97	44.33	-40.5	-41.75	0.01	NO
miR-107	13	4449289	4449309	+	833	AGCAGCATTGTACAGGGCTAT	21	3	72	48.61	-34.4	-47.78	0.01	NO
miR-146	13	7555658	7555676	-	112	TGAGAACTGAATTCCATGG	19	4	72	44.44	-32.1	-44.58	0.01	NO
miR-365	14	764324	764343	+	18	TAATGCCCCTAAAAATCCTT	20	2	138	47.1	-54	-39.13	0.01	NO
N/A	14	857078	857098	+	26	GCGGAAGGACGGCGTCACTGG	21	1	66	72.73	-36.2	-54.85	0.01	NO
N/A	14	4018538	4018560	+	111	CAGCAGGACTGGCTTTGTTACGA	23	1	84	59.52	-38.9	-46.31	0.01	NO
miR-454	15	399864	399882	-	47	TAGTGCAATATTGCTTATA	19	1	89	41.57	-39.7	-44.61	0.01	NO
miR-301	15	406327	406349	-	424	CAGTGCAATAATATTGTCAAAGC	23	1	85	42.35	-29.34	-34.52	0.01	NO
miR-130	15	408409	408430	-	1639	CAGTGCAATATTAAAAGGGCAT	22	1	109	44.04	-37.9	-34.77	0.01	NO
N/A	15	769601	769618	+	25	ATCCCTTACTCACATGAG	18	1	88	50	-53.1	-60.34	0.01	NO
miR-1306	15	1296957	1296978	+	65	TGGACGTTGGCTCTGGTGGTGA	22	1	82	57.32	-33.7	-41.1	0.01	NO
N/A	15	10171741	10171758	+	12	CTGGAGGACACAGAGGCA	18	2	139	64.03	-67	-48.2	0.01	NO
miR-455	17	5339754	5339774	+	90	GCAGTCCATGGGCATATACAC	21	1	106	50.94	-44	-41.51	0.01	TATGTGCCCTTGGACTACATC
miR-199	17	5667203	5667224	+	149	ACAGTAGTCTGCACATTGGTTA	22	2	69	49.28	-29.6	-42.9	0.01	NO
miR-126	17	8431793	8431812	-	37	CATTATTACTTTTGGTACGC	20	1	85	49.41	-41.3	-48.59	0.01	CGTACCGTGAGTAATAATGC
miR-181	17	10218559	10218580	+	203	ACCATCGACCGTTGACTGTACC	22	1	95	41.05	-32.7	-34.42	0.01	AACATTCAACGCTGTCGGTGAG
miR-181	17	10220152	10220175	+	294	AACATTCATTGCTGTCGGTGGGTT	24	2	120	50.83	-48.1	-40.08	0.01	NO
miR-365	18	6437352	6437371	+	18	TAATGCCCCTAAAAATCCTT	20	2	83	42.17	-35	-42.17	0.01	NO
N/A	18	10554165	10554185	-	29	CGGCTTCTCGGTACCTGCGTT	21	1	73	58.9	-31.76	-43.51	0.01	NO
miR-142	19	497036	497057	-	25	CCCATAAAGTAGAAAGCACTAC	22	1	86	43.02	-43.6	-50.7	0.01	NO
miR-22	19	5352156	5352175	-	33	AGTTCTTCAGTGGCAAGCTT	20	1	90	48.89	-45.6	-50.67	0.01	NO
miR-144	19	5824134	5824154	-	76	TACAGTATAGATGATGTACTC	21	1	76	46.05	-34.1	-44.87	0.01	NO
N/A	19	7145042	7145061	-	199	CAGTGCAATGTTAAAAGGGC	20	1	92	46.74	-41.3	-44.89	0.01	NO
miR-21	19	7322089	7322111	+	223	TAGCTTATCAGACTGATGTTGAC	23	1	90	45.56	-45.4	-50.44	0.01	NO
miR-1	20	8107876	8107897	+	694	TGGAATGTAAAGAAGTATGTAT	22	2	83	39.76	-39.1	-47.11	0.01	NO
miR-133	20	8119113	8119134	+	45	TTTGGTCCCCTTCAACCAGCTG	22	3	84	41.67	-34	-40.48	0.01	NO
miR-130	20	8681835	8681852	+	181	TAAGGCACGCGGTGAATG	18	2	76	50	-34.5	-45.39	0.01	NO
miR-200	21	2583333	2583352	-	32	TAACACTGTCTGGTAACGAT	20	1	135	46.67	-53.5	-39.63	0.01	NO
miR-200	21	2585663	2585681	-	33	TAATACTGCCTGGTAATGA	19	1	84	41.67	-36.82	-43.83	0.01	NO
N/A	22	2685020	2685041	-	29	AAGGTCCAACCTCACATGTCCT	22	1	161	46.58	-77.64	-48.22	0.01	NO
N/A	23	1213487	1213505	+	13	TGCGCTTTCTCATCCCGGC	19	1	61	63.93	-31.7	-51.97	0.01	NO
miR-133	23	4664098	4664119	+	45	TTTGGTCCCCTTCAACCAGCTG	22	3	81	54.32	-42.3	-52.22	0.01	NO
miR-30	23	5248474	5248495	+	40045	CTTTCAGTCGGATGTTTACAGC	22	1	88	55.68	-47.9	-54.43	0.01	TGTAAACATCCTTGACTGGAAG
miR-30	23	5249653	5249675	+	45	TGTAAACATCCTACACTCTCAGC	23	2	87	52.87	-37	-42.53	0.01	NO
miR-100	24	3372906	3372927	+	173	AACCCGTAGATCCGAACTTGTG	22	1	63	42.86	-23.3	-36.98	0.01	NO
let-7	24	3380997	3381018	+	95	TGAGGTAGTAGGTTGTATAGTT	22	4	138	47.1	-54.2	-39.28	0.01	NO
let-7	26	1442757	1442778	-	95	TGAGGTAGTAGGTTGTATAGTT	22	4	103	45.63	-52.3	-50.78	0.01	NO
let-7	26	1442959	1442977	-	26	CTGAGGTAGTAGATTGAAT	19	1	98	43.88	-39.7	-40.51	0.01	NO
miR-29	26	2512579	2512598	-	29	TAGCACCATTTGAAATCAGT	20	2	113	43.36	-39.46	-34.92	0.01	NO
miR-205	26	2896068	2896089	+	26	TCCTTCATTCCACCGGAGTCTG	22	1	93	48.39	-41.62	-44.75	0.01	NO
miR-196	27	3553120	3553141	+	29	TAGGTAGTTTCATGTTGTTGGG	22	3	86	33.72	-32.2	-37.44	0.01	NO
miR-9	28	2709415	2709436	+	35	TAAAGCTAGAGAACCGAATGTA	22	1	84	39.29	-32.3	-38.45	0.01	NO
N/A	MT	9050	9068	+	31	AGCTAGAGAGAGGGGACAC	19	1	107	45.79	-23.6	-22.06	0.01	NO
miR-10	Un_random	379349	379370	-	28660	TACCCTGTAGATCCGAATTTGT	22	1	86	38.37	-34.76	-40.42	0.01	NO
miR-146	Un_random	14731567	14731588	+	1596	TGAGAACTGAATTCCATGGACT	22	2	116	45.69	-52.43	-45.2	0.01	NO
N/A	Un_random	16238906	16238925	+	31	ATGGGTCAAACGTTGACCAA	20	3	109	47.71	-44.5	-40.83	0.01	NO
N/A	Un_random	21596967	21596984	-	17	GCAGGAGCGGGGCTCGGT	18	11	66	75.76	-41.4	-62.73	0.01	NO
miR-196	Un_random	27776519	27776540	-	29	TAGGTAGTTTCATGTTGTTGGG	22	3	84	41.67	-37.8	-45	0.01	NO
N/A	Un_random	38326406	38326429	+	20	TCCCAGTGGAGCTCTGCAAGGACC	24	1	65	66.15	-33.9	-52.15	0.01	NO
N/A	Un_random	47015331	47015348	-	17	GCAGGAGCGGGGCTCGGT	18	11	66	75.76	-41.4	-62.73	0.01	NO
N/A	Un_random	61344147	61344164	-	17	GCAGGAGCGGGGCTCGGT	18	11	66	75.76	-41.4	-62.73	0.01	NO
N/A	Un_random	63782874	63782891	+	17	GCAGGAGCGGGGCTCGGT	18	11	66	75.76	-41.4	-62.73	0.01	NO
miR-101	Z	28037921	28037942	+	233	GTACAGTACTGTGATAACTGAA	22	2	88	42.05	-37.2	-42.27	0.01	NO
N/A	Z	34596495	34596514	+	31	TCCTTAACTCATGCCGCTGT	20	1	82	35.37	-31	-37.8	0.01	NO
miR-23	Z	41157420	41157440	+	32	GGGTTCCTGGCATGATGATTT	21	1	84	42.86	-39.1	-46.55	0.01	NO
miR-27	Z	41157702	41157722	+	58	TTCACAGTGGCTAAGTTCTGC	21	1	77	46.75	-41.2	-53.51	0.01	NO
miR-24	Z	41158218	41158238	+	26	TGGCTCAGTTCAGCAGGAACA	21	1	75	50.67	-28.5	-38	0.01	NO

Continued on Next Page...

miR	Chr	Start	End	Ori	Abun	Seq	Len	G. Hits	H. Len	G/C%	MFE	AMFE	p-value	miRNA*
N/A	Z	44167547	44167568	-	1825	AAAGGACGGAGGCGGCCCGCGC	22	1	81	74.07	-51.3	-63.33	0.01	NO
miR-9	Z	59286330	59286352	+	34	TCTTTGGTTATCTAGCTGTATGA	23	2	85	36.47	-38.4	-45.18	0.01	NO
N/A	Z	68816780	68816803	-	841	ATGCAGAAGTGCACGGAAACAGCT	24	1	89	49.44	-40.4	-45.39	0.01	NO
miR-31	Z	71882219	71882240	-	1317	AGGCAAGATGTTGGCATAGCTG	22	1	109	45.87	-49.5	-45.41	0.01	NO

Appendix B

B.1 Conserved fruit miRNAs

B.2 Conserved leaf miRNAs

B.3 Conserved tomato miRNAs

Table B.1: Cloning frequency of conserved tomato miRNAs from fruit.

miRNA	Exact	Shorter	Longer	1 mismatch	2 mismatches	Total
miR156	3	0	0	0	1	4
miR159	207	94	86	15	17	419
miR160	29	4	11	0	2	46
miR162	37	0	43	7	4	91
miR164	23	4	6	1	0	34
miR165	0	0	2	0	0	2
miR166	261	34	243	10	55	603
miR167	271	75	52	3	31	432
miR168	32	312	4	5878	2339	8565
miR169	427	36	145	30	28	666
miR170	18	3	2	0	2	25
miR171	2075	365	516	137	165	3258
miR172	61	30	36	4	10	141
miR319	19	0	7	0	4	30
miR390	25	15	5	0	5	50
miR393	2	0	0	0	0	2
miR394	0	0	0	0	0	0
miR395	0	0	1	0	0	1
miR396	49	28	10	9	18	114
miR397	0	0	0	0	0	0
miR398	0	0	0	4	4	8
miR399	2	0	0	4	0	6
miR403	0	0	0	0	20	20
miR408	0	0	0	0	3	3
miR472	0	0	0	0	0	0
miR482	0	0	0	0	7	7
miR828	0	0	0	0	6	6
miR858	0	0	0	0	92	92
miR894	0	0	2	1	1	4
miR1151	0	0	0	0	0	0

Table showing exact matches, “shorter” (where the sRNA is a substring of the miRNA), “longer” (where the miRNA is a substring of the sRNA) and sRNAs with one and two mismatches to the miRNA along with the total numbers.

Table B.2: Cloning frequency of conserved tomato miRNAs from leaf.

miRNA	Exact	Shorter	Longer	1 mismatch	2 mismatches	Total
miR156	47	27	7	6	2	89
miR159	156	247	43	6	28	480
miR160	19	3	9	7	1	39
miR162	141	9	16	19	15	200
miR164	80	10	17	8	8	123
miR165	2	0	1	0	1	4
miR166	142	17	23	12	34	228
miR167	210	115	41	119	50	535
miR168	41	390	6	7913	2092	10442
miR169	19	1	7	0	0	27
miR170	10	5	1	0	2	18
miR171	3388	1270	281	350	278	5567
miR172	21	51	13	1	9	95
miR319	7	1	2	1	1	12
miR390	0	3	0	1	11	15
miR393	1	0	1	0	0	2
miR394	9	0	9	2	2	22
miR395	0	0	0	0	1	1
miR396	73	99	23	18	57	270
miR397	0	0	0	0	2	2
miR398	0	0	0	3	24	27
miR399	0	0	0	6	1	7
miR403	0	0	0	0	74	74
miR408	5	23	3	0	8	39
miR472	0	0	0	1	1	2
miR482	0	0	0	0	31	31
miR828	0	0	0	0	0	0
miR858	0	0	0	5	360	365
miR894	0	0	0	0	1	1
miR1151	0	0	0	0	2	2

Table showing exact matches, “shorter” (where the sRNA is a substring of the miRNA), “longer” (where the miRNA is a substring of the sRNA) and sRNAs with one and two mismatches to the miRNA along with the total numbers.

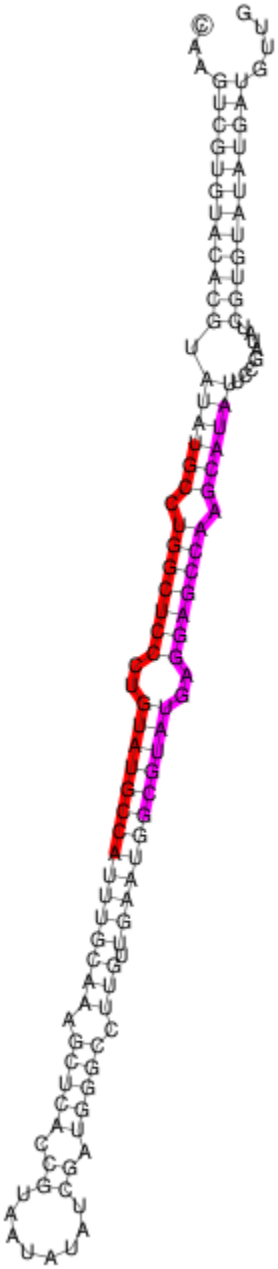

Figure B.1: Secondary structure of sly-miR160. miRNA is highlighted in red and miRNA* in pink.

Figure B.2: Secondary structure of sly-miR167. miRNA is highlighted in red.

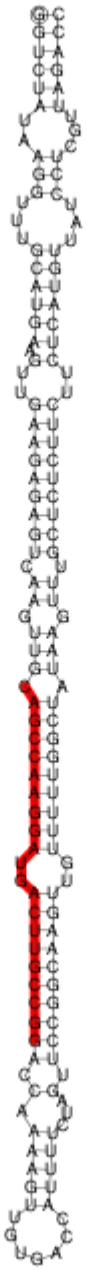

Figure B.3: Secondary structure of sly-miR169a. miRNA is highlighted in red.

Figure B.4: Secondary structure of sly-miR169b. miRNA is highlighted in red.

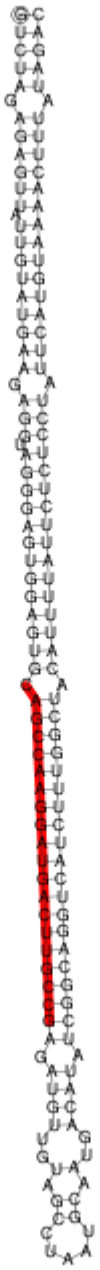

Figure B.5: Secondary structure of sly-miR169c. miRNA is highlighted in red.

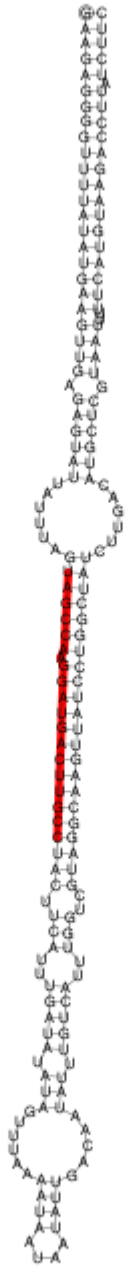

Figure B.6: Secondary structure of sly-miR169d. miRNA is highlighted in red.

Figure B.7: Secondary structure of sly-miR171a. miRNA is highlighted in red and miRNA* in pink.

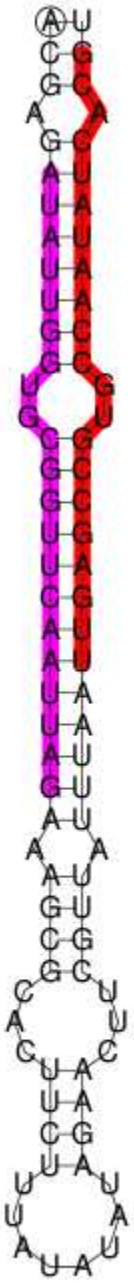

Figure B.8: Secondary structure of sly-miR171b. miRNA is highlighted in red and miRNA* in pink.

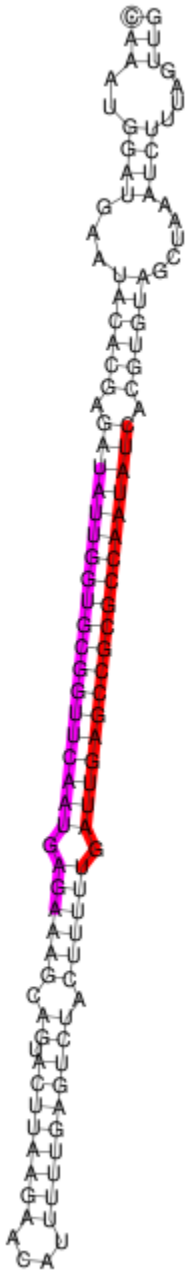

Figure B.9: Secondary structure of sly-miR171c. miRNA is highlighted in red and miRNA* in pink.

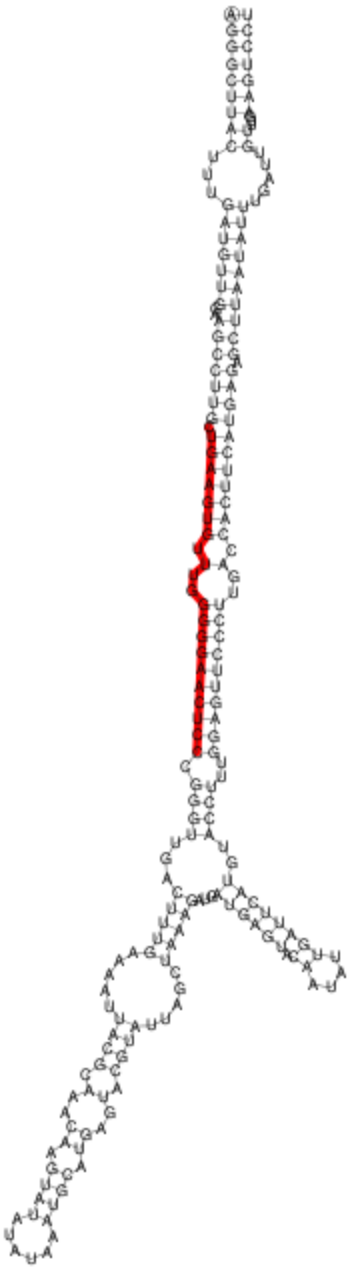

Figure B.10: Secondary structure of sly-miR395a. miRNA is highlighted in red.

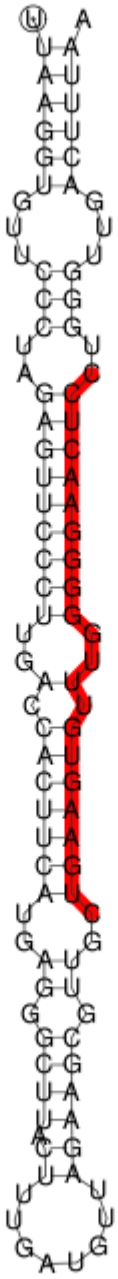

Figure B.11: Secondary structure of sly-miR395b. miRNA is highlighted in red.

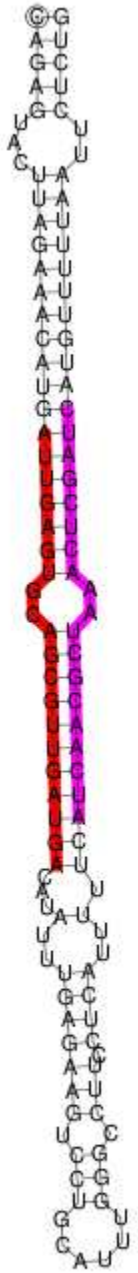

Figure B.12: Secondary structure of sly-miR397. miRNA is highlighted in red and miRNA* in pink.

B.4 Novel tomato miRNAs

Figure B.13: Secondary structure of sly-miRW. miRNA is highlighted in red.

Figure B.14: Secondary structure of sly-miRX. miRNA is highlighted in red.

Figure B.15: Secondary structure of sly-miRY. miRNA is highlighted in red.

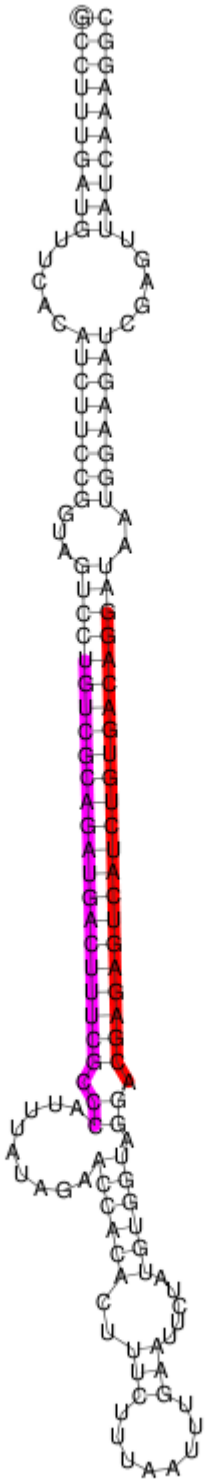

Figure B.16: Secondary structure of sly-miRZ. miRNA is highlighted in red and miRNA* in pink.

Appendix C

C.1 Feature selection

Listed below are the 39 features used to train the SVM in Chapter 6.

Sequence Length

In plants there is a strong bias towards 21nt miRNAs and miRNA* sequences, with the majority of plant miRNAs in the central miRNA repository, miRBase [Griffiths-Jones et al., 2008] falling into this size class (Figure. C.1).

Features chosen:

- Length of predicted miRNA
- Length of predicted miRNA*
- Difference in length between the miRNA and miRNA*

Alignment properties and binding energy

The miRNA/miRNA* duplex are excised from characteristic miRNA precursor hairpins (Figure. 2.1) and often show a typical 3' two nucleotide overhang (Figure. C.2). They are not usually perfectly complementary to one another and may contain several mismatches and in some cases can contain "bulges" or asymmetrical unpaired bases.

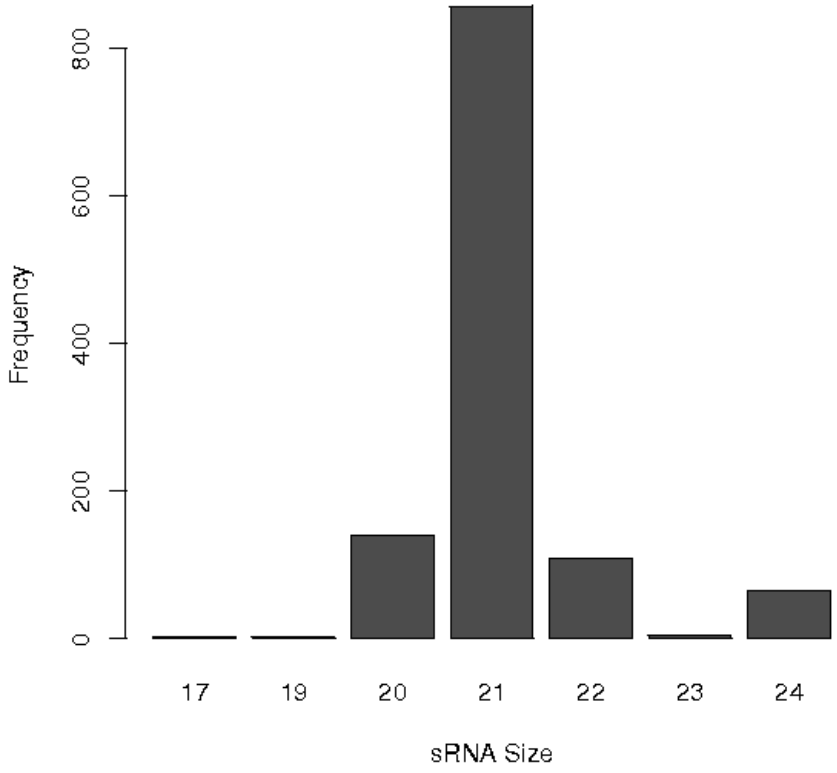

Figure C.1: Histogram showing the frequency of mature plant miRNA size classes in miRBase 11.0 (April 2008)

in addition miRNA/miRNA* duplexes are energetically stable and MFE is an important feature in distinguishing miRNA/miRNA* pairs.

Features chosen:

- Complementarity score (+1 for canonical base pair between miRNA and miRNA* and +0.5 for a G-U base)
- Number of mismatches between miRNA and miRNA* sequence
- Number of non-canonical (G-U) base pairs

```
ath-miR161        5′ -            C         - 3′
                      UGAAAGUGACUA AUCGGGGU
                      |||||||||||| |:|:||
                    UCACUUUCACUGAU UGGUCC
ath-miR161*     3′ -              U       - 5′
```

Figure C.2: Complementary miRNA/miRNA* duplex of *Arabidopsis thaliana ath-miR161* and *ath-miR161** showing the characteristic 2nt 3' overhang

- Number of asymmetrical unpaired bases (bulges)
- Length of miRNA 3' overhang
- Length of miRNA* 3' overhang
- Length of miRNA 5' overhang
- Length of miRNA* 5' overhang
- MFE of the miRNA/miRNA* duplex
- Adjusted free energy (AMFE) of the miRNA/miRNA* duplex (MFE per 100nt)

Nucleotide composition

It is known that plant miRNAs show a bias towards uridine bases at the first position in the sequence [Mi et al., 2008]. For this reason we capture the first four bases of both the miRNA and miRNA* and use these as features in the SVM. Nucleotide composition of both miRNA and miRNA* sequences can also be useful in distinguishing real miRNA/miRNA* duplexes and can be effective in filtering out low-complexity sequences.

Features chosen:

- miRNA percentage A composition
- miRNA percentage G composition
- miRNA percentage C composition
- miRNA percentage U composition
- miRNA* percentage A composition
- miRNA* percentage G composition
- miRNA* percentage C composition
- miRNA* percentage U composition
- First base of the miRNA
- Second base of the miRNA
- Third base of the miRNA
- Fourth base of the miRNA
- First base of the miRNA*
- Second base of the miRNA*
- Third base of the miRNA*
- Fourth base of the miRNA*

Base pairing properties

miRNA/miRNA* base-pairing (complementarity) is often imperfect including multiple mismatches or non-canonical G-U base pairing. Several base pairing features can be used in order to capture information about the complementarity between miRNA and miRNA*.

Features chosen:

- Total number of unpaired bases in the miRNA
- Maximum number of consecutive unpaired bases in the miRNA
- Total number of unpaired bases in the miRNA*
- Maximum number of consecutive unpaired bases in the miRNA*
- Number of consecutive unpaired 5’ nucleotides in predicted miRNA
- Number of consecutive unpaired 3’ nucleotides in predicted miRNA
- Number of internal unpaired nucleotides in predicted miRNA
- Number of consecutive unpaired 5’ nucleotides in predicted miRNA*
- Number of consecutive unpaired 3’ nucleotides in predicted miRNA*
- Number of internal unpaired nucleotides in predicted miRNA*

Appendix D

D.1 No genome miRNA prediction results

D.1.1 *Arabidopsis* results

Column “Predicted miRNA” shows the sequence of the predicted mature miRNA; column “Predicted miRNA*” shows the sequence of the predicted mature miRNA*; column “p-value” shows the p-value assigned by LIBSVM to this prediction; column “Correct locus?” shows whether the predicted miRNA/miRNA* are present at the same genomic locus; column “miRNA?” shows whether the predicted miRNA sequence is in fact a *bona-fide* miRNA sequence.

Table D.1: miRNAs predicted in *Arabidopsis* using the no genome SVM method with a p-value threshold of 0.90

miRNA	miRNA*	p-value	Correct locus?	miRNA?
TGCGGGAAGCATTTGCACATGT	ACATGTGCAAATGCTTTCTACA	1.0000	Yes	Yes
AGAATCTTGATGATGCTGCA	GCAGCACCATTAAGATTCAC	1.0000	Yes	Yes
AAAGCTCAGGAGGGATAGCGCC	TGGCGCTATCCATCCTGAGTT	1.0000	Yes	Yes
TTGAGCCGTGCCAATATCAC	AGATATTGGTGCGGTTCAATC	1.0000	Yes	Yes
TGAAAGTGACTACATCGGGG	ACCCTGGTTTAGTCACTTTCA	1.0000	Yes	Yes
TGGAGAAGCAGGGCACGTGCG	GCACGTGTTCTACTACTCCAAC	1.0000	Yes	Yes
TTGGACTGAAGGGAGCTCCTT	GGAGATTCTTTCAGTCCAGTC	1.0000	Yes	Yes
TTCGCTTGCAGAGAGAAATCAC	TGATTCTCTGTGTAAGCGAAA	1.0000	Yes	Yes
TTCCACAGCTTTCTTGAACTT	GCTCAAGAAAGCTGTGGGAAA	1.0000	Yes	Yes
TGCGGGAAGCATTTGCACATG	CATGTGCAAATGCTTTCTACA	1.0000	Yes	Yes
GCTCTCTATACTTCTGTCACC	TTGACAGAAGATAGAGAGCA	1.0000	Yes	Yes
TGAAGCTGCCAGCATGATCTA	TAGATCATGTTCGCAGTTTCA	1.0000	Yes	Yes
TGCAGCCAAGGATGACTTGCC	GCAAGTTGTCCTTGGCTACA	1.0000	Yes	Yes
TGAAAGTGACTACATCGGGGTT	ACCCTGGTTTAGTCACTTTCA	1.0000	Yes	Yes
TTGAGCCGTGCCAATATCACG	AGATATTGGTGCGGTTCAATC	1.0000	Yes	Yes
TCAATGGTGTCTAATAAGTTTT	TAACTTATTAGACACCATGAT	1.0000	No	No
TTCTAAGTCCAACATAGCATA	CGCTATGTTGGACTTAGAATA	1.0000	No	No
TAACTTATTAGACACCATGA	CAATGGTGTCTAATAAGTTT	1.0000	No	No

Continued on Next Page. . .

miRNA	miRNA*	p-value	Correct locus?	miRNA?
TCCTAAGTCCAACATAGCGTT	TGCTATGTTGGACTTAGAATA	1.0000	No	No
TTCTACCATCCGATCAACAAG	TGATTGATAGGATGGTAGAAG	1.0000	No	No
GGGCATTTTCGTGATTTGTG	GCCCAAATCACGAAAATGCCC	1.0000	No	No
TTTTGCATATCCTGGAATATG	ACATATTTCAGTATATGCAAA	1.0000	No	No
GGGCATTTTCGTGATTTGTGC	GCCCAAATCACGAAAATGCCC	1.0000	No	No
TTGATGATTCGACAAAGTGAA	TCACTTTGTCGAGTCACCAAG	1.0000	Yes	Yes
TGAAGCTGCCAGCATGATCTAA	TAGATCATGTTCGCAGTTTCA	1.0000	Yes	Yes
GATCATGTTCGCAGTTTCACC	GAAGCTGCCAGCATGATCTA	1.0000	Yes	Yes
ATATTCCAGGATATGCAAAAG	TTTTGCATATCCTAGAATATA	1.0000	No	No
TATTCCAGGATATGCAAAAGA	TTTTGCATATCCTAGAATATA	1.0000	No	No
GGGCATTTTCGTGATTTGGGC	GCACAAATCACGAAAATGCCC	1.0000	No	No
TCAATGGTGTCTAATAAGTTT	TAACTTATTAGACACCATGAT	1.0000	No	No
TCGCTTGGTGCAGGTCGGGAA	TCCCGCCTTGCATCAACTGAA	1.0000	Yes	Yes
TCAATGCATTGAAAGTGACTA	TTTAGTCACTTTCACTGCATT	1.0000	Yes	Yes
TAGACCATTTGTGAGAAGGGA	CCCTTCTCATCGATGGTCTAG	1.0000	Yes	Yes
TTGACAGAAGATAGAGAGCAC	GCTCTCTATACTTCTGTCACC	1.0000	Yes	Yes
CAGCCAAGGATGACTTGCCGA	GGCAAGTTGTCCTTGGCTAC	1.0000	Yes	Yes
GTTCAATAAAGCTGTGGGAAG	TTCCACAGCTTTCTTGAACT	1.0000	Yes	Yes
TTTAGTCACTTTCACTGCATT	CAATGCATTGAAAGTGACTA	1.0000	Yes	Yes
GCTCAAGAAAGCTGTGGGAAA	TTCCACAGCTTTCTTGAACT	1.0000	Yes	Yes
TAGGACGCATAACATACTGGTA	TACCAGTATCTTATGCGTCCTA	1.0000	No	No
AGATAGGACGCATAACATACTG	CAGTATCTTATGCGTCCTATCT	1.0000	No	No
AAGCTCAGGAGGGATAGCGCC	TGGCGCTATCCATCCTGAGTT	1.0000	Yes	Yes
TTGAAAGTGACTACATCGGG	CCTGGTTTAGTCACTTTCACT	1.0000	Yes	Yes
AAGCTCAGGAGGGATAGCGC	GCGCTATCCATCCTGAGTTCC	1.0000	Yes	Yes
TTGAAAGTGACTACATCGGGG	ACCCTGGTTTAGTCACTTTCA	1.0000	Yes	Yes
CGGTTCAATAAAGCTGTGGGA	TTCCACAGCTTTCTTGAACT	1.0000	Yes	Yes
GAATCTTGATGATGCTGCAT	GCAGCACCATTAAGATTCAC	1.0000	Yes	Yes
TGAAGCTGCCAGCATGATCT	TAGATCATGTTCGCAGTTTCA	1.0000	Yes	Yes
TTGAAAGTGACTACATCGGGGT	ACCCTGGTTTAGTCACTTTCA	1.0000	Yes	Yes
TGAAAGTGACTACATCGGGGT	ACCCTGGTTTAGTCACTTTCA	1.0000	Yes	Yes
CAGCCAAGGATGACTTGCCG	GCAAGTTGTCCTTGGCTACA	1.0000	Yes	Yes
TGACAGAAGAGAGTGAGCAC	GCTCACTGCTCTTTCTGTCAG	1.0000	Yes	Yes
TCGCTTGGTGCAGGTCGGGA	TCCCGCCTTGCATCAACTGAA	1.0000	Yes	Yes
TTCCACAGCTTTCTTGAACTG	CGGTTCAATAAAGCTGTGGGA	1.0000	Yes	Yes
CCTTCTCATCGATGGTCTAGA	TAGACCATTTGTGAGAAGGG	1.0000	Yes	Yes
TGGCGCTATCCATCCTGAGTT	AAGCTCAGGAGGGATAGCGCCA	1.0000	Yes	Yes
TGATGATTCGACAAAGTGAAG	TCACTTTGTCGAGTCACCAAG	1.0000	Yes	Yes
TGGAGAAGCAGGGCACGTGC	GCACGTGTTCTACTACTCCAAC	1.0000	Yes	Yes
GCAAGTTGACCTTGGCTCTGC	TGAGCCAAGGATGACTTGCC	1.0000	Yes	Yes
GAAGCTGCCAGCATGATCTA	TAGATCATGTTCGCAGTTTCA	1.0000	Yes	Yes
GGCAAGTTGTCCTTGGCTACA	GGTAGCCAAGGATGACTTGCC	1.0000	No	Yes
TGATTGAGCCGCGCCAATATC	AGATATTAGTGCGGTTCAATC	1.0000	No	Yes
GTGGCATCATCAAGATTCAC	AGAATCTTGATGATGCTGCAC	1.0000	No	Yes
TGAGCCAAGGATGACTTGCCG	GGCAAGTTGTCCTTGGCTAC	1.0000	No	Yes
AGATATTGGTGCGGTTCAATC	TGATTGAGCCGCGCCAATATCC	1.0000	No	Yes
GGAATCTTGATGATGCTGCAT	GGAGCATCATCAAGATTCACA	1.0000	No	Yes
TTGGACTGAAGGGAGCTCCC	GGAGATTCTTTCAGTCCAGTC	1.0000	No	Yes
TTGGACTGAAGGGAGCTCCCT	GGAGATTCTTTCAGTCCAGTC	1.0000	No	Yes
GGAGATTCTTTCAGTCCAGTC	GATTGGACTGAAGGGAGCTCC	1.0000	No	Yes
CTTGGACTGAAGGGAGCTCCC	GGAGATTCTTTCAGTCCAGTC	1.0000	No	Yes
CGACAGAAGAGAGTGAGCAC	GCTCACTCTCTTTTTGTCATA	1.0000	No	Yes
TGGAGAAGCAGGGCACGTGCAT	GCACGTGTTCTACTACTCCAAC	1.0000	No	Yes
TGAAGCTGCCAGCATGATCTAT	TAGATCATGTTCGCAGTTTCA	1.0000	No	Yes
TTGACAGAAGAGAGTGAGCAC	GCTTACTCTCTCTCTGTCACC	1.0000	No	Yes
TGGAGAAGCAGGGCACGTGCA	GCACGTGTTCTACTACTCCAAC	1.0000	No	Yes
ATGAAGCTGCCAGCATGATCTA	TAGGTCATGCTGGTAGTTTCAC	1.0000	No	Yes
TGAAGCTGCCAGCATGATCTG	TAGATCATGTTCGCAGTTTCA	1.0000	No	Yes
GGCAAGTTGTCCTTGGCTAC	TGAGCCAAGGATGACTTGCC	1.0000	No	Yes
TGAAGCTGCCAGCATGATCTGG	TAGATCATGTTCGCAGTTTCA	1.0000	No	Yes
GTGAAGCTGCCAGCATGATCTA	TAGATCATGTTCGCAGTTTCA	1.0000	No	Yes
GGGCATCTTTCTATTGGCAGG	CCTGCCAAAGGAGAGTTGCCC	1.0000	No	Yes
TCGCTTGGTGCAGGTCGGGAAC	TCCCGCCTTGCATCAACTGAA	0.9900	Yes	Yes
AGAGCTTCCTTGAGTCCATTC	GATTGGACTGAAGGGAGCTCC	0.9900	Yes	Yes

Continued on Next Page. . .

miRNA	miRNA*	p-value	Correct locus?	miRNA?
TGATTCTCTGTGTAAGCGAAA	TTCGCTTGCAGAGAGAAATCA	0.9900	Yes	Yes
CGCTATCCATCCTGAGTTTC	AAGCTCAGGAGGGATAGCGC	0.9900	Yes	Yes
CATGTGCAAATGCTTTCTACAG	TGCGGGAAGCATTTGCACATGT	0.9900	Yes	Yes
CGCTATCCATCCTGAGTTCC	AAGCTCAGGAGGGATAGCGC	0.9900	Yes	Yes
TAAGCTGCCAGCATGATCTTG	TAGGTCATGCTGGTAGTTTCAC	0.9900	Yes	Yes
TGCCTGGCTCCCTGTATGCCA	GCGTACAAGGAGTCAAGCATG	0.9900	Yes	Yes
TCCCAAATGTAGACAAAGCA	CTTTGTCTACAATTTTGGAA	0.9900	Yes	Yes
TGCCAAAGGAGAGTTGCCCTG	GGGCATCTTTCTATTGGCAGG	0.9900	Yes	Yes
TGACAGAAGAGAGTGAGCACA	GCTCACTGCTCTTTCTGTCAG	0.9900	Yes	Yes
TTGCCGACCCTCAGTAGGAGC	TGCGTTCCTACCGAGGTCGGC	0.9900	No	No
GCGGAGGACATTGTCAGGTGG	TCACTTGATGATGTTCTTCGA	0.9900	No	No
AGATATTAGTGCGGTTCAATC	TGATTGAGCCGCGCCAATATCT	0.9900	No	Yes
GAGAATCTTGATGATGCTGCAT	GGAGCATCATCAAGATTCACA	0.9900	No	Yes
CAGCCAAGGATGACTTGCCGG	GCAAGTTGTCCTTGGCTACA	0.9900	No	Yes
TAGCCAAGGATGACTTGCCTG	TGGCAAGTTGTCCTTCGGCTAC	0.9900	No	Yes
GGCAAGTTGTCCTTCGGCTACA	GGTAGCCAAGGATGACTTGCC	0.9900	No	Yes
AGAATCTTGATGATGCTGCAT	GGAGCATCATCAAGATTCACA	0.9900	No	Yes
TAGCCAAGGATGACTTGCCTGT	TGGCAAGTTGTCCTTCGGCTAC	0.9900	No	Yes
TGTTGATCGGATGGTAGAAAC	TTCTTCTACCATCCTATCAAT	0.9900	No	No
TTTGGATTGAAGGGAGCTCT	AGAGCTTCCTTGAGTCCATTC	0.9900	No	Yes
ATGCCTGGCTCCCTGTATGCC	GCGTACAAGGAGTCAAGCATG	0.9900	Yes	Yes
TACGAGCCACTTGAAACTGAA	TCAATTTCTAGTGGGTCGTAT	0.9900	Yes	Yes
GGCAAGTTGTCCTTCGGCTAC	TGAGCCAAGGATGACTTGCC	0.9900	No	Yes
GGAGAAGCAGGGCACGTGCA	GCACGTGTTCTACTACTCCAAC	0.9900	No	Yes
TTTGGATTGAAGGGAGCTCTA	AGAGCTTCCTTGAGTCCATTC	0.9900	No	Yes
TGTGTTCTCAGGTCACCCCTG	AGGGTTGATATGAGAACACAC	0.9900	Yes	Yes
GGAGGGCGCGGCGGTCGCTG	GCCCGCTGCGCCTCCTCCGA	0.9900	No	No
AGAATCTTGATGATGCTGCAG	GGAGCATCATCAAGATTCACA	0.9900	Yes	Yes
TGTGTTCTCAGGTCACCCCTT	AGGGTTGATATGAGAACACAC	0.9900	No	Yes
TGCCTGGCTCCCTGTATGCC	GCGTATGAGGAGCCATGCAT	0.9900	Yes	Yes
TTGACAGAAGAAAGAGAGCAC	GCTCTCTTTCCTTCTGCCACC	0.9900	Yes	Yes
CGCTATCCATCCTGAGTTTCA	AAAGCTCAGGAGGGATAGCGC	0.9900	Yes	Yes
GAACTAGAAAAGACATTGGAC	TATCCAATGCTTTTTCTAGTT	0.9900	No	No
TGGCAGAGTGGCCTTGCTGCC	TGGTGCAAGGTATACTTTGTT	0.9900	No	No
GCTCTCTAGCCTTCTGTCATC	TGACAGAAGATAGAGAGCAC	0.9900	Yes	Yes
GCAAGTTGACCTTGGCTCTGT	TGAGCCAAGGATGACTTGCC	0.9900	Yes	Yes
CATGTGCAAATGCTTTCTACA	CGGGAAGCATTTGCACATGTT	0.9900	Yes	Yes
TTGTTGATCGGATGGTAGAAA	TTCTTCTACCATCCTATCAAT	0.9900	No	No
TAAGCTGCCAGCATGATCTTGT	TAGGTCATGCTGGTAGTTTCAC	0.9900	Yes	Yes
GCAGCACCATTAAGATTCACA	TGAGAATCTTGATGATGCTGC	0.9900	Yes	Yes
TTCTAAGTTCAACATATCGAC	CGCTATGTTGGACTTAGAATA	0.9900	No	No
TCAATGCATTGAAAGTGACT	TCACTTTCACTGCATTAATC	0.9900	Yes	Yes
TCGAGTTCCAACCTCTTCAAC	TTGAAGAGGACTTGGAACTTC	0.9900	Yes	Yes
TTTCTACCATCCGATCAACAAG	TGATTGATAGGATGGTAGAAG	0.9900	No	No
TTGGATTGAAGGGAGCTCTA	AGAGCTTCCTTGAGTCCATTC	0.9900	No	No
TGAGAATCTTGATGATGCTGC	GGAGCATCATCAAGATTCACA	0.9900	Yes	Yes
ATGGAGAAGCAGGGCACGTGCA	GCACGTGTTCTACTACTCCAAC	0.9800	No	Yes
TATGAGAGTATTATAAGTCAC	TATTTGTAATATTTTTATGTT	0.9800	No	Yes
TTGTTGATCGGATGGTAGAAAC	TTCTTCTACCATCCTATCAAT	0.9800	No	No
GGGTTGATATGAGAACACACG	TGTGTTCTCAGGTCACCCCT	0.9800	Yes	Yes
TTCGGACCAGGCTTCATTCCC	TGAATGATGCCTGGCTCGAGA	0.9800	No	Yes
GGGCAACTGCGTGGTAACTGGT	CCTGTTACCACTGAAGTTGCCC	0.9800	No	No
GCTCTCTATACTTCTGTCACCA	TGACAGAAGAAAGAGAGCAC	0.9800	No	Yes
CGCGGATTACGGTGGCGGCCT	TGCCGTGATCGTGGTCTGCA	0.9800	Yes	Yes
CCTGGTTTAGTCACTTTCACT	GAAAGTGACTACATCGGGGT	0.9800	Yes	Yes
AGAATCTTGATGATGCTGCATT	GGAGCATCATCAAGATTCACA	0.9800	No	Yes
TTAGATTCACGCACAAACTC	TGTTTTGTGCTTGAATCTAA	0.9800	Yes	Yes
TATTGGCCTGGTTCACTCAGA	TGATTGAGCCGCGCCAATAT	0.9800	Yes	Yes
TTGGCATTCTGTCCACCTCC	AGGTGGGCATACTGCCAATA	0.9800	Yes	Yes
TCGGACCAGGCTTCATTCCCC	TGAATGATGCCTGGCTCGAGA	0.9800	Yes	Yes
TCGGACCAGGCTTCATTCCC	TGAATGATGCCTGGCTCGAGA	0.9800	Yes	Yes
GCGTACAAGGAGTCAAGCATG	TGCCTGGCTCCCTGTATGCC	0.9800	Yes	Yes
ATTGAAAGTGACTACATCGGGG	ACCCTGGTTTAGTCACTTTCA	0.9800	Yes	Yes
TTACTTATGGTACCGTAGTAA	TACTTATGGTACCGTAGTAAA	0.9800	Yes	Yes

Continued on Next Page. . .

miRNA	miRNA*	p-value	Correct locus?	miRNA?
TCCAATGCTTTTTCTAGTTTCG	GAACTAGAAAAGACATTGGA	0.9800	No	No
TGTGCAAATGCTTTCTACAGG	TGCGGGAAGCATTTGCACATGT	0.9800	Yes	Yes
TGCGAGAAGCATTTGCACATG	ACATGTGCAAATGCTTTCTAC	0.9800	No	No
TTTGGATTGAAGGGAGCTCTT	AGAGCTTCCTTGAGTCCATTC	0.9800	No	Yes
AGAATCTTGATGATGCTGCATC	GGAGCATCATCAAGATTCACA	0.9800	No	Yes
TGTTGATCGGATGGTAGAAACA	TTCTTCTACCATCCTATCAAT	0.9800	No	No
GCAGCACCATTAAGATTCAC	TGGGAATCTTGATGATGCTGC	0.9800	Yes	Yes
TTTGGATTGAAGGGAGCTCTAC	AGAGCTTCCTTGAGTCCATTC	0.9700	No	Yes
TGACAGAAGAAAGAGAGCAC	GCTCTCTAGCCTTCTGTCATCA	0.9700	No	Yes
TTAGATGACCATCAACAAACT	TAGTTTGTTTGATGGTAACTA	0.9700	No	Yes
GCTCACTGCTCTATCTGTCAGA	CGACAGAAGAGAGTGAGCAC	0.9700	No	Yes
CGGACCAGGCTTCATCCCCC	GGATGGGTCGGCCGGTCCGC	0.9700	No	Yes
GGCCTCGATGAGTAGGAGGGC	CCCTTCTCATCGATGGTCTAG	0.9700	No	No
TTGAATTGAAGTGCTTGAATT	TTCAAGGACTTCTATTCAGA	0.9700	Yes	Yes
TTCGCAGGAGAGATAGCGCCA	TGGCGCTATCCATCCTGAGTT	0.9700	No	Yes
GTGTTCTCAGGTCACCCCTGC	AGGGTTGATATGAGAACACAC	0.9700	Yes	Yes
AGAGCTTTCTTCGGTCCACTC	GATTGGACTGAAGGGAGCTCC	0.9700	No	Yes
TTAGATTCACGCACAAACTCG	TGTTTTGTGCTTGAATCTAAT	0.9700	Yes	Yes
GGAGGCAGCGGTTCATCGATC	TCGATAAACCTCTGCATCCA	0.9700	Yes	Yes
AGCTTCCTTGAGTCCATTCAC	GATTGGACTGAAGGGAGCTCC	0.9700	Yes	Yes
CGCTATCCATCCTGAGTTCCA	AAAGCTCAGGAGGGATAGCGC	0.9700	Yes	Yes
TCGATAAACCTCTGCATCCAG	GGAGGCAGCGGTTCATCGATC	0.9700	Yes	Yes
TTTTGCATATACTCGAATACC	TATTCTAGGATATGCAAAAGT	0.9700	No	No
TCCAATGTCTTTTCTAGTTCG	AAACTAGAAAAAGCATTGGATA	0.9700	No	No
TGGGCAACTGCGTGGTAACTGG	CCTGTTACCACTGAAGTTGCCC	0.9700	No	No
ACATGTGCAAATGCTTTCTACA	GCGGGAAGCATTTGCACATGT	0.9700	Yes	Yes
GCAGCACCATTAAGATTCACAT	GGAATCTTGATGATGCTGCA	0.9700	No	Yes
TCCCAAATGTAGACAAAGCAA	TCTTTGTCTACAATTTTGGAAA	0.9600	Yes	Yes
ATGTGCAAATGCTTTCTACAG	TGCGGGAAGCATTTGCACATGT	0.9600	Yes	Yes
CCTATACCCGGCCGTCGGGGC	CCAAATGGCTGGGTTTTGGGA	0.9600	No	No
TTCTAAGTCCAACATAGCGTA	CGATATGTTGAACTTAGAATA	0.9600	No	No
TCCAATGTCTTTTCTAGTTCGT	AAACTAGAAAAAGCATTGGATA	0.9600	No	No
GGTAGGACGTGTCGGCTGCT	CAGTATCTTATGCGTCCTATCT	0.9600	No	No
TGTGTAATTGTGTGTCAGCCA	TGGGTGACACATCATCACACA	0.9600	No	No
TCTCAAGAAGGTGCATGAACA	TACTATGCTGCCATCTTGAGAT	0.9500	No	Yes
TTCGGACCAGGCTTCATTCCCC	TGAATGATGCCTGGCTCGAGA	0.9500	No	Yes
GACGCGGATTACGGTGGCGGCC	TGCCGTGATCGTGGTCTGCA	0.9500	Yes	Yes
TTTGCATATACTCGAATACCT	ACATATTTCAGTATATGCAAA	0.9500	No	No
TGTGAACATATTCAAGGATAAC	TATTGTTTTGAATGTGTTCAA	0.9500	No	No
AGAATCCTGATGATGCTGCAT	GGAGCATCATCAAGATTCACA	0.9500	No	No
TCCCGCCTTGCATCAACTGAA	CGCTTGGTGCAGGTCGGGAA	0.9500	Yes	Yes
TCGGACCAGGCTTCATCCCC	GGGATGGGTCGGCCGGTCCGC	0.9500	No	Yes
GCGTATGAGGAGCCATGCATA	TGCCTGGCTCCCTGTATGCC	0.9500	Yes	Yes
TGGAGGCAGCGGTTCATCGATC	CGATAAACCTCTGCATCCAG	0.9400	Yes	Yes
TCAATGCATTGAAAGTGACTAC	GTCACTTTCACTGCATTAATC	0.9400	Yes	Yes
TTGCATATACTCGAATACCTA	ACATATTTCAGTATATGCAAA	0.9400	No	No
CGGACCAGGCTTCATTCCCC	TGAATGATGCCTGGCTCGAGA	0.9400	Yes	Yes
GTTATTCTATTCCACCTCTTA	TTAGGAGGTTGAATGAGTAGT	0.9400	No	No
TCGGACCAGGCTTCATTCCCCC	TGAATGATGCCTGGCTCGAGA	0.9400	No	Yes
CCCTACTGATGCCCGCGTCGCG	TACGCGACGGGGTATTGTAAG	0.9400	Yes	Yes
CTGACAGAAGATAGAGAGCAC	GCTCACTGCTCTTTCTGTCAG	0.9400	No	Yes
CGGAGCTGTGTCAACTCGTGC	GCAAGTTGACTTTGGCTCTGT	0.9400	No	No
TATTCTAGGATATGCAAAAGT	TGTTTTCTTTTGCATATCCTGG	0.9300	No	No
CCCGCCTTGCATCAACTGAAT	TCGCTTGGTGCAGGTCGGGA	0.9300	Yes	Yes
TGCCACGATCCACTGAGATTC	TACTCTCAGAGGATCAGTTGC	0.9300	No	No
CCTAACGGCGTGCCTCGGCATC	GCCGAGGGCACGTCTGCCTGGG	0.9300	Yes	Yes
TGCACTGCCTCTTCCCTGGC	GGAGAAGCAGGGCACGTGCAT	0.9300	No	Yes
GCACATGGGTTAGTCGATCC	GGATCGCATGGCCTCTGTGCT	0.9300	No	No
CGTACAAGGAGTCAAGCATGA	TGCCTGGCTCCCTGTATGCC	0.9300	Yes	Yes
CCAAAACCCGGTGGATAAAA	TTATCCCCCGTGTTTTGTCC	0.9300	Yes	Yes
GGCAAGTGTGCCACGAGACCGG	TCGGATTTTGGATACTTGCCT	0.9300	No	No
CGATCCCCCGCATCTCCACCA	GGTGGAGGTGTTGGGTTCGAGA	0.9300	No	No
TTTATCTAGATGATGCATTTC	TATTGTGTTTTATCTAGATGA	0.9300	Yes	Yes
TAGCCAAGGATGACTTGCCTGA	GCAAGTTGACCTTGGCTCTGT	0.9200	No	Yes

Continued on Next Page. . .

miRNA	miRNA*	p-value	Correct locus?	miRNA?
CGGAGGACATTGTCAGGTGG	CACGTGGCAATGAACTCCTTC	0.9200	No	No
GATCCCCGGCAACGGCGCCA	GAGGTGTTGTCTTCGTGGATCT	0.9200	No	No
TGATTGAGCCGTGTCAATATC	AGATATTAGTGCGGTTCAATC	0.9200	No	Yes
TGCGCCCGCCGCCCGATTGCC	GAGTATTCGATTGCGGCGGCGC	0.9200	No	No
GGTCGGCTTGTCCCTTCGGTC	TACCACAGGGATAACTGGCTT	0.9200	No	No
TCATTGAGTGCATCGTTGATG	TGTGGATGATGCACTCAATCT	0.9200	No	Yes
TGGAGTGATGCTTCTCGACTA	TACGTGGAGGCATCCCTTCAC	0.9200	No	No
CATTCAAGGACTTCTATTCAG	TTTGAATTGAAGTGCTTGAAT	0.9200	Yes	Yes
CCGACGCGGATTACGGTGGCG	TCGCTGCCGTGATCGTGGTCT	0.9200	Yes	Yes
TAATACTAAACATATTCATGG	TCGAATATGTTTTGTATTATT	0.9200	No	No
CGACGCGGATTACGGTGGCGGC	TGCCGTGATCGTGGTCTGCA	0.9200	Yes	Yes
CCCTACTGATGCCCGCGTCGC	TACGCGACGGGGTATTGTAAG	0.9200	Yes	Yes
TCTTTGTCTACAATTTTGGAAA	TTCCCAAATGTAGACAAAGCA	0.9200	Yes	Yes
TGTGTTTTATCTAGATGATGC	TGTTTTATCTAGATGATGCAT	0.9200	Yes	Yes
TCGGACCAGGCTTCATCCCCC	GGATGGGTCGGCCGGTCCGC	0.9100	No	Yes
CACTGAAGGACCTAAACTAAC	TTGTTTAGGTCCCTTAGTTT	0.9100	Yes	Yes
TTAGATTCACGCACAAACTCGT	TGTTTTGTGCTTGAATCTAAT	0.9100	Yes	Yes
TGGTCGGCTTGTCCCTTCGGT	TACCACAGGGATAACTGGCTT	0.9100	No	No
CGCTATGTTGGACTTAGGATG	TTCTAAGTTCAACATATCGACG	0.9100	No	No
TCGATCCCCGGCAACGGCGCCA	GAGGTGTTGTCTTCGTGGATCT	0.9100	No	No
CAATGCATTGAAAGTGACTA	GTCACTTTCACTGCATTAATC	0.9000	Yes	Yes
TTTGGATTGAAGGGAGCTCTTC	GGAGATTCTTTCAGTCCAGTC	0.9000	No	Yes
CGACGGGGTATTGTAAGTGGC	TGGTCACAACAATATTCCGTTG	0.9000	No	No
CGGAGGACATTGTCAGGTGGG	CACGTGGCAATGAACTCCTTC	0.9000	No	No
TGACAGAAGATAGAGAGCAC	TCTCTAGCCTTCTGTCATCACC	0.9000	Yes	Yes
TGCACTGCCTCTTCCCTGGCT	GGAGAAGCAGGGCACGTGCAT	0.9000	No	Yes

D.1.2 *Oryza sativa* results

Column “Predicted miRNA” shows the sequence of the predicted mature miRNA; column “Predicted miRNA*” shows the sequence of the predicted mature miRNA*; column “p-value” shows the p-value assigned by LIBSVM to this prediction; column “Correct locus?” shows whether the predicted miRNA/miRNA* are present at the same genomic locus; column “miRNA?” shows whether the predicted miRNA sequence is in fact a *bona-fide* miRNA sequence.

Table D.2: miRNAs predicted in rice using the no genome SVM method with a p-value threshold of 0.90

miRNA	miRNA*	p-value	Correct locus?	miRNA?
CAGCCAAGGATGACTTGCCGG	GGCAAGTCTGTCCTTGGCTAC	0.9989	Yes	Yes
CAGCCAAGGATGACTTGCCGA	CGGCAAGTTGTCCTTGGCTAC	0.9986	No	Yes
TCGCTTGGTGCAGATCGGGA	CCCGCCTTGCACCAAGTGAA	0.9985	Yes	Yes
TGGAAGGGGCATGCAGAGGAG	CCTGTGCTTGCCTCTTCCAT	0.9985	Yes	Yes
TCGCTTGGTGCAGATCGGGAC	TCCCGCCTTGCACCAAGTGAAT	0.9981	Yes	Yes
TCGCTTGGTGCAGATCGGGACC	TCCCGCCTTGCACCAAGTGAAT	0.9981	Yes	Yes
TTGGACTGAAGGGTGCTCCCT	GAGCTCCTTTCGGTCCAAAAA	0.9980	No	Yes
TGCACTGCCTCTTCCCTGGC	CAGGGATGAGGCAGAGCATGG	0.9972	Yes	Yes
TGACAGAAGAGAGTGAGCAC	GCTCACTTCTCTTTCTGTCAG	0.9971	Yes	Yes
CGACGCGCCGCGGCCGGCCGC	CCGGCACGGTGGCTGCGCGTC	0.9962	No	No
GGCAGTCTCCTTGGCTAGCC	ATAGCCAAGGATGACTTGCCT	0.9950	Yes	Yes
TGATTGAGCCGCGCCAATATC	TGTTGGCATGGTTCAATCAAA	0.9949	No	Yes
TTGACAGAAGAGAGTGAGCAC	GCTCACTCCTCTTTCTGTCACC	0.9944	Yes	Yes
CCCGCCTTGCACCAAGTGAAT	TCGCTTGGTGCAGATCGGGA	0.9941	Yes	Yes
TTGGATTGAAGGGAGCTCTG	GAGCGTCCTTCAGTCCACTC	0.9914	No	Yes
CTGCACTGCCTCTTCCCTGGC	CAGGGATGAGGCAGAGCATGG	0.9912	Yes	Yes
TGCCTGGCTCCCTGTATGCCG	GCATTGAGGGAGTCATGCAGG	0.9818	No	Yes
TAGCCAAGAATGACTTGCCTA	CGGCAAGTTGTCCTTGGCTAC	0.9771	No	Yes
GCGCGCGACCCACACCAGGCC	TACCTGGTGTGAATTGCA	0.9713	No	No
CCCCGGCAGAGAGCGCGACC	GGACGTGTTCCGGCTGCCGG	0.9684	No	No
AGGGCCCGTGCCACCGGCCAA	GCCGGAGGTAGGGTCCAGC	0.9680	No	No
CGGCAAGCTAGAGACAGCAAC	TGCAGTTGTTGTCTCAAGCTTG	0.9677	No	Yes
ATGCCTGGCTCCCTGTATGCCA	GCGTGCGAGGAGCCAAGCATGA	0.9646	Yes	Yes
TGGTGAGCCTTCCTGGCTAAG	TTAGCCAAGAATGGCTTGCCTA	0.9641	Yes	Yes
GCTGACGAGCGGGAGGCCCT	GGGGGCCTTCCCCGGGCGT	0.9623	No	No
TCCACAGGCTTTCTTGAACTG	TTCAAGAAAGTCCTTGGAAA	0.9613	Yes	Yes
TTGAGTGCAGCGTTGATGAAC	TCACCAGCACTGCACCCAATC	0.9556	Yes	Yes
CCCGCCACCGGCGCCGCTTCC	GCGGTGGCGGAGGTGGGGGCTG	0.9550	No	No
CCCCGCGTCGCACGGATTCGT	CAGAACTGGCGATGCGGGAT	0.9453	No	No
TCGGACCAGGCTTCATTCCCC	GGGCGATGAATCAGGTCCGAC	0.9417	No	Yes
TGCCTGGCTCCCTGAATGCCA	GCGTGCGAGGAGCCAAGCATG	0.9364	No	Yes
TCGGACCAGGCTTCATTCCC	TGAACCGGAAGCCTGGTT	0.9164	No	Yes
TCGGACCAGGCTTCATTCCTC	GAACCGGAAGCCTGGTTA	0.9155	No	Yes
TGCCAAAGGAGAGTTGCCCTG	GGCAGTCTCCTTGGCTAG	0.9139	No	Yes
TGGCCGCTGATGACCCACCTC	GGGGTCTCCGGTGGCCACGGC	0.9028	No	No
TGTGGTCTTGCCTATGTGGCA	TCCACATGGCGTGCCACGTAA	0.9024	No	No
TTGAGTGCAGCGTTGATGAACC	TCACCAGCACTGCACCCAATC	0.9014	Yes	Yes

D.1.3 *Solanum lycopersicum* results

Column “Predicted miRNA” shows the sequence of the predicted mature miRNA; column “Predicted miRNA*” shows the sequence of the predicted mature miRNA*; column “p-value” shows the p-value assigned by `LIBSVM` to this prediction; column “miRNA abundance” shows the abundance of the predicted miRNA; column “miRNA* abundance” shows the abundance of the predicted miRNA*.

Table D.3: miRNAs predicted in tomato using the no genome SVM method with a p-value threshold of 0.90

miRNA	miRNA*	p-value	miRNA abundance	miRNA* abundance
CAGACACGTTGTCCTAACCGA	TCGGTTAGGACAACATGTCTG	0.9999	26	17
TAGGACAACATGTCTGGATGAC	TCATCCAGACACGTTGTCCTAA	0.9999	80	14
TCGGTTAGGACAACATGTCTGG	CCAGACACGTTGTCCTAACCGA	0.9999	197	67
CAGACACGTTGTCCTAACCGAC	TCGGTTAGGACAACATGTCTGT	0.9999	107	17
TTAGGACAACATGTCTGGATGA	TCATCCAGACACGTTGTCCTAA	0.9999	92	67
TTGAGCCGTGCCAATATCAC	TGATGTTGGAATGGCTCAAT	0.9998	70	8
CGGTTAGGACAACATGTCTGG	CCAGACACGTTGTCCTAACCG	0.9998	47	8
TTTATCGTGAGTGGCACATGGT	AACCATGTGCCACTCTCGATAA	0.9997	26	13
TCCATTTACGTGCAAGCGCAG	CTGCGCTTGCCCGTAAATGGA	0.9997	27	16
ACATATATAGTCACACCGATTA	TAATCGGTGTGACTATATATGC	0.9997	20	15
TCCATTTACGTGCAAGCGCAGT	ACTGCGCTTGCCCGTAAATGGA	0.9997	45	16
CGGTTAGGACAACATGTCTGGA	CCAGACACGTTGTCCTAACCGA	0.9996	250	67
CCAGACACGTTGTCCTAACCGA	TCGGTTAGGACAACATGTCTG	0.9996	67	5
TCAACCATGTGCCACTCTCGA	TCGTGAGTGGCACATGGTTAAT	0.9996	36	7
TCAACCATGTGCCACTCTCGAT	TCGTGAGTGGCACATGGTTAAT	0.9995	50	26
TTGAGCCGTGCCAATATCACG	TGATGTTGGAATGGCTCAAT	0.9994	96	8
GTGTGCTGGATTATGACTGAA	GCTCAGTCATAATCCAGCACA	0.9994	29	5
TTCCATGAGACTGTTTTTGGGT	TCCCAAAAACGGTCTTATGGA	0.9994	34	32
ACATGCGAGTATGCCTCATGTA	TACATGAGGCATACCCGCATGT	0.9994	162	5
ATGTCTGGATGACACAGGTGCC	CACCTGTGTCATCCAGACACGT	0.9993	53	5
TTGGCTGAGTGAGCATCACGG	TCAGGTGCTCACTCAGCTAAT	0.9992	259	11
TACCGTGTGCTGGATTATGACA	AGTCATAATCCAGCACACGG	0.9991	38	7
TTGGCTGAGTGAGCATCACTG	GAGGTGCTCACTCAGCTAATA	0.9990	80	66
CTTGGGACCAAAGTCACCAA	TGGTGACTTTGATCTCAAAA	0.9990	26	14
TGCGAGTATGCCTCATGTACC	TACATGAGGCATACCCGCATGT	0.9990	23	12
ATGCGAGTATGCCTCATGTACC	TACATGAGGCATACCCGCATGT	0.9989	70	12
ACATGCGAGTATGCCTCATGT	TACATGAGGCATACCCGCATGT	0.9988	40	6
GTTCAAGAAAGTTGTGGGAA	ATTCCACAGCTTTCTTGAAC	0.9988	23	6
CTTTTAAGGACCTATCAAGAAT	TTCTTGGTAGGTCCTTAAAAAT	0.9987	22	11
CAGGTGCTCACTCAGCTAAT	TTGGCTGAGTGAGCATCACT	0.9987	20	10
TGCGAGTATGCCTCATGTACCA	TACATGAGGCATACCCGCATGT	0.9986	140	12
TGATTGAGCCGCGCCAATATC	TATTGGTGCGGTTCAATGAG	0.9986	38	7
AATCTGTTTGATCACTTACAAA	TTTGTAAGTGATCAAACAGATA	0.9984	42	6
CGTGTGCTGGATTATGACTGAA	GCTCAGTCATAATCCAGCACA	0.9983	77	5
GTTCAATAAAGCTGTGGGAAG	TTCCACAGCTTTCTTGAACT	0.9982	73	11
TTCCACAGCTTTCTTGAACTG	CGGTTCAATAAAGCTGTGGGA	0.9981	99	20
TATTGGTGCGGTTCAATGAGA	TGATTGAGCCGCGCCAATAT	0.9980	50	7
CGTGTGCTGGATTATGACTGAT	GCTCAGTCATAATCCAGCACA	0.9980	27	5
GAGGTGCTCACTCAGCTAATA	TTGGCTGAGTGAGCATCACT	0.9979	47	20
CACAGGTGCCACTTATTTATG	CCATAAATAAGTGGCACATGTG	0.9978	78	10
TTGGCTGAGTGAGCATCACGGA	TCAGGTGCTCACTCAGCTAAT	0.9978	20	20
TGGAAGGGAGAATATCCAGGA	CTGGATATTATCCTTTCCATC	0.9977	88	16

Continued on Next Page. . .

miRNA	miRNA*	p-value	miRNA abundance	miRNA* abundance
CACAGGTGCCACTTATTTATGA	CCATAAATAAGTGGCACATGTG	0.9977	60	13
TGGATGACACAGGTGCCACTTA	AATAAGTGGCACATGTGTCATC	0.9976	135	13
TGGTATTGTTCCGTTCAGGGA	CCTGAACGGAACAATACGAT	0.9974	33	24
CTTGGGACCAAAGTCACCAAC	TGGTGACTTTGATCTCAAAA	0.9974	1349	14
AAAGGTTTCGTTTGAAAGCTTT	CAAGCTTTCAAACGAAACCTTT	0.9969	24	14
CTTGGGACCAAAGTCACCAAT	TGGTGACTTTGATCTCAAAA	0.9965	23	18
ATGACACAGGTGCCACTTATT	AATAAGTGGCACATGTGTCATC	0.9965	46	13
CACGGTAGCTTCGCGCCACTGG	ATGGCGCGAAGCTACCGTGTGC	0.9964	28	7
CAGGTGCTCACTCAGCTAATA	TTGGCTGAGTGAGCATCACT	0.9964	110	10
TGACACAGGTGCCACTTATT	AATAAGTGGCACATGTGTCATC	0.9964	41	10
GTTCAAGAAAGTTGTGGGAAA	ATTCCACAGCTTTCTTGAAC	0.9964	22	6
CGTTTGTGCGTGAATCTAAC	CTAGATTCACGCACAAGCTC	0.9962	60	23
CCCGCCTTGCATCAACTGAAT	TCGCTTGGTGCAGGGCGGGAC	0.9959	27	12
GACACAGGTGCCACTTATTTA	ATAAATAAGTGGCACATGTGTC	0.9958	64	13
CGGTGATAATGGTATTCTAA	TTGGAATACCATCATCACCG	0.9958	53	17
CAGATTACTTTTGCTGGTACA	TGTACCAACGAAAGTAATCTG	0.9957	36	12
ATGACACAGGTGCCACTTATTT	AATAAGTGGCACATGTGTCATC	0.9953	34	10
TCGGACCAGGCTTCATTCCCC	GAATGTTGTCTGGTTCGAAAA	0.9953	179	10
TCTCGGACCAGGCTTCATTCC	GAATGTTGTCTGGTTCGAAAA	0.9952	20	10
TGACACAGGTGCCACTTATTT	AATAAGTGGCACATGTGTCATC	0.9951	38	10
TCGCTTGGTGCAGGCCGGGAC	TCCCGGCCGAGCATGAGGTGC	0.9948	47	36
TACTTTTGCTGGTACATGAGGC	TGCCTCATGTACCAACGAAAG	0.9945	40	5
TGACACAGGTGCCACTTATTTA	ATAAATAAGTGGCACATGTGTC	0.9944	283	13
TGATTGAGCCGTGCCAATATC	TGATGTTGGAATGGCTCAAT	0.9943	5217	50
TGGACGCCCATGAGGTACTGG	TACCAGTGCCCAGGGTGTCTA	0.9942	50	14
GCGCTCCGGACGCTGGCCTG	GCGTGGTGTCCGGTGCGCTC	0.9940	32	14
ACACAGGTGCCACTTATTTATG	CCATAAATAAGTGGCACATGTG	0.9939	270	10
CGGTTCAATAAAGCTGTGGGA	TTCCACAGCTTTCTTGAACTT	0.9939	20	10
TTAGGTACACGATCTAGGAATA	CATATTCTTAGATCGTGTACCC	0.9937	20	8
TGATTGAGCCGTGCCAATAA	TATTGGTGCGGTTCAATGAG	0.9937	120	6
CGGTGATAATGGTATTCTAAT	TTGGAATACCATCATCACCGT	0.9937	49	17
ACGGTGATAATGGTATTCTAA	TTGGAATACCATCATCACCGT	0.9937	413	17
TCGGACCAGGCTTCATTCCTC	GAATGTTGTCTGGTTCGAAAA	0.9936	196	10
CAGATTACTTTTGCTGGTACAT	TGTACCAACGAAAGTAATCTGT	0.9934	33	12
TGAAGCTGCCAGCATGATCT	GATCATGTGGTTGCTTCACC	0.9930	44	19
CAACTTTGTCATTTAGAGCTGA	CAGCTCTAAATAACAAAGTTG	0.9928	20	5
GGATGACACAGGTGCCACTTA	AATAAGTGGCACATGTGTCATC	0.9928	27	13
ACGGTGATAATGGTATTCTA	TTGGAATACCATCATCACCGT	0.9922	67	17
TGATTGAGCCGTGCCAATAT	TATTGGTGCGGTTCAATGAG	0.9922	464	50
CTTGGGACCAAAGTCACCAACA	TGGTGACTTTGATCTCAAAA	0.9920	50	14
TGATTGAGCCGTGCCAAAAA	TATTGGTGCGGTTCAATTAG	0.9919	24	13
TATTGGCCTGGTTCACTCAGA	TGATTGAGCCATGCCAATATC	0.9918	210	120
TGATTGAGCCGTGCCAATATCA	TGATGTTGGAATGGCTCAAT	0.9918	426	8
TGAAGCTGCCAGCATGATCTA	TCAGATCATGTGGTTGCTTCA	0.9917	380	20
TTACTTTTGCTGGTACATGAG	CTCATGTACCAACGAAAGTAA	0.9911	29	14
ACGGTGATAATGGTATTCTAAA	TTGGAATACCATCATCACCGT	0.9909	38	17
CTTGGGACCAAAGTCACCAACT	TGGTGACTTTGATCTCAAAA	0.9909	31	14
CTAGATTCACGCACAAGCTCG	CGTTTGTGCGTGAATCTAAC	0.9905	93	60
CTGGATTATGACTGAACGCC	GCTCAGTCATAATCCAGCACA	0.9905	23	5
TTACTTTTGCTGGTACATGAGG	CTCATGTACCAACGAAAGTAAT	0.9903	51	14
CGTGTCCCACCGGGTGTGCCA	TAAACACCCGGTGCGGATAA	0.9902	55	17
TGAAGCTGCCAGCATGATCTAA	TCAGATCATGTGGTTGCTTCA	0.9901	85	19
TATCCAAAGACAATCCATGGAA	TTCCATGAGACTGTTTTTGGGT	0.9901	52	34
ATTACTTTTGCTGGTACATGAG	CTCATGTACCAACGAAAGTAAT	0.9896	42	14
ACGGTGATAATGGTATTCTAAT	TTGGAATACCATCATCACCGT	0.9890	20	17
TCTTTCCTACTCCTCCCATACC	TGTGGGTGGGGTGGAAAGATT	0.9886	35	33
CACTTTGTTGGTGACTTTGAT	CCAAAGTCACCAACAGAGTGTC	0.9874	29	7
CCGTGTGCTGGATTATGACTGT	GCTCAGTCATAATCCAGCACA	0.9871	22	5
CTGGATTATGACTGAACGCCT	GCTCAGTCATAATCCAGCACA	0.9868	28	5
AACGGTGATAATGGTATTCTA	TTGGAATACCATCATCACCGT	0.9864	29	17
TGATTGAGCCGTGCCAATACC	TATTGGTGCGGTTCAATGAG	0.9864	35	5
ATGAAGCTGCCAGCATGATCTA	TCAGATCATGTGGTTGCTTCA	0.9863	21	19
TGGATTATGACTGAACGCCTC	GCTCAGTCATAATCCAGCACA	0.9861	99	47
CGGGTCCCACCGGGTGTGCCA	CAGGTCTCGGTGGGACCTCCA	0.9860	35	16

Continued on Next Page. . .

miRNA	miRNA*	p-value	miRNA abundance	miRNA* abundance
ATGATTGAGCCGTGCCAATATC	TATTGGTGCGGTTCAATTAG	0.9856	103	13
TCGGACCAGGCTTCATTCCCCA	GAATGTTGTCTGGTTCGAAAA	0.9853	110	10
TGATTGAGCCGTGCCAATATCT	TGATGTTGGAATGGCTCAAT	0.9852	113	8
TGAAGCTGCCAGCATGATCTAT	TCAGATCATGTGGTTGCTTCA	0.9848	27	19
TGGAGAAGCAGGGCACGTGCA	TATGTGTCCTGTTTTCATAAT	0.9847	101	6
TGATTGAGCCGTGCCAATAGC	TATTGGTGCGGTTCAATGAG	0.9847	41	33
CCACAAAGGCCTTTGGTGGA	CCACAAAGGCTTTTGGTGGA	0.9845	23	5
TGCCTGGCTCCCTGTATGCCA	GCGTATGAGGAGCCAAGCATA	0.9845	47	23
CACGATCTAGGAATATGTTGTA	CAACATATTCCTGGACCGTGTA	0.9842	45	8
TCTCTTGGTGCAGGTCGGGAC	TGCCGACCTGCAGTAGGGGCC	0.9842	20	8
TGATTGAGCCGTGCCAATATT	TGATGTTGGAATGGCTCAAT	0.9840	67	50
TCGGGTCCCACCGGGTGTGCCA	CAGGTCTCGGTGGGACCTCCA	0.9829	23	16
TGATTGAGCCGTGCCAATAAA	TATTGGTGCGGTTCAATGAG	0.9827	43	8
TCCTTGGTGCAGGTCGGGAC	TGCCGACCTGCAGTAGGGGCC	0.9821	35	18
ACAGGTGCCACTTATTTATGA	CCATAAATAAGTGGCACATGTG	0.9821	57	13
CGAGTATGCCTCATGTACCAAC	TTGCTGGTACATGAGGCATACC	0.9815	110	40
TCGCTTGGTGCAGGTCGGGA	CCCGCCTTGCATCAACTGAAT	0.9814	641	27
TCGGACCAGGCTTCATTCCTCA	GAATGTTGTCTGGTTCGAAAA	0.9805	64	10
TTCAAAGTTTCCGACGGGTGC	CACTTTGTTGGTGACTTTGAA	0.9804	21	5
TCGCTTGGTGCAGGTCGGGGC	CCCGCCTTGCATCAACTGAAT	0.9799	131	27
TCGCTTGGTGCAGGTCGGGAC	CCCGCCTTGCATCAACTGAAT	0.9795	13559	56
TGATTGAGCCGTGCCAATATA	TATTGGTGCGGTTCAATGAG	0.9789	141	50
TCGCTTGGTGCGGGTCGGGAC	TCCCGGCCGAGCATGAGGTGC	0.9789	21	9
ATGTGCCACTCTCGATAAATTC	TTATCGTGAGTGGCACATGGT	0.9786	59	7
CAGTGACCATGACAACTTCA	GATGTTGTCATGGTAATTGTC	0.9775	20	7
TTTACGTGCAAGCGCAGTTGAA	ACTGCGCTTGCCCGTAAATGGA	0.9773	28	16
ACAGGTGCCACTTATTTATGAA	CCATAAATAAGTGGCACATGTG	0.9766	846	13
ACTTGGGACCAAAGTCACCAAC	TGGTGACTTTGATCTCAAAA	0.9765	40	14
CTGGATGACACAGGTGCCACTT	AATAAGTGGCACATGTGTCATC	0.9760	64	10
TAAAGCTGCCAGCATGATCTGG	TCAGATCATGTGGTTGCTTCA	0.9760	117	13
TATTGGCCTGGTTCACTCAGAA	TGATTGAGCCATGCCAATATC	0.9759	33	8
TCGCTTGGTGCAGGTCGGGAA	CCCGCCTTGCATCAACTGAAT	0.9758	64	27
AACGGTGATAATGGTATTCTAA	TTGGAATACCATCATCACCGT	0.9749	38	17
TCGCTTGGTGCAGGTCGGGC	CCTGCCTTGCATCAACTGAA	0.9744	26	9
TCTGGATGACACAGGTGCCACT	AATAAGTGGCACATGTGTCATC	0.9742	740	20
TCTTGCCTACACCGCTCATGCC	CGTGAGCGGTGGGGAAAGATA	0.9727	20	11
CAAACAGATTACTTTTGCTGGT	CAACGAAAGTAATCTGTTTGAT	0.9726	29	12
GGGTTACGGTGCCAAACTGC	TTGGCATGGTAGCCCTATAA	0.9724	36	15
TCGCTTGGTGCAGGTCGGGAT	CCCGCCTTGCATCAACTGAAT	0.9723	175	27
CTTAAAAGCTGGTCAAACTGAC	AAGTCAGTTTAATCAGCTTTT	0.9717	20	5
TTTGGATGACACAGGTGCCACT	AATAAGTGGCACATGTGTCATC	0.9715	22	10
TTCATGGGCCAACAAAGAAGAT	TCTTTACCGTTGGTCCATGGAA	0.9695	20	6
CGGGTCCCACTGGGCGTGCCA	TGATGCCTACTGGGTACCTGT	0.9694	79	16
CTTGGGACCAAAAGTCACCAAC	TGGTGACTTTGATCTCAAAA	0.9656	58	12
TCGCTTGGTGCAGGTCAGGAC	CCTGCCTTGCATCAACTGAA	0.9649	20	8
TGGCGCGAAGCTACCGTGTGC	TTACACGGTAGTAAGGTGCTA	0.9640	76	6
CACGACCTCTCACTGGTGCTGC	AGCACCACTGAGAGGTAGTGCT	0.9639	23	5
CCGCTTGGTGCAGGTCGGGAC	TCCCGGCCGAGCATGAGGTGC	0.9637	38	27
ACACAGGTGCCACTTATTTA	CCATAAATAAGTGGCACATGTG	0.9634	30	13
TATGTAGGGCCCATATGAGGAC	CATTTCATGTGGGCCTAACATC	0.9630	23	7
TTTCTCAGGTGCTCACTCAGC	TTGGCTGAGTGAGCATCACT	0.9610	66	6
TCGGTGCAGATCTTGGTGGT	TCCGCCCTAGGTGTGCACCGG	0.9609	34	15
TCGTGTCCCACCGGGTGTGCCA	TAAACACCCGGTGCGGATAA	0.9609	26	16
CTGGATGACACAGGTGCCACT	AATAAGTGGCACATGTGTCATC	0.9602	28	10
TGATTGAGCCGTGTCAATATC	TGATGTTGGAATGGCTCAAT	0.9593	28	8
CGTCGCGGTGACCGCCTTGAA	TTGGCTAAGTCGCCGCGACGG	0.9593	28	16
TGGATTATGACTGAACGCCTCA	GCTCAGTCATAATCCAGCACA	0.9592	45	5
GCTCCGGACGCTGGCCTGTG	TATGGGTGCTCCGGTGCATG	0.9590	68	15
TCGGGTCCCACTGGGCGTGCCA	CAGGTCTCGGTGGGACCTCCA	0.9590	27	17
TTGACAAGGTGGGTAATCTGG	TAGATTATGTGCCTCGTTAAA	0.9583	103	5
GTCGGTGCAGATCTTGGTGGT	TCCGCCCTAGGTGTGCACCGGC	0.9578	28	15
CGCGCTCCGGACGCTGGCCTG	GCGTGGTGTCCGGTGCGCTC	0.9577	63	14
TTTCCGCCGGGTGCGCATTGGG	CGGCGATGCGCCCCGGTCGGA	0.9551	20	9
TCGCTTGGTGCAGGTCGGGACC	TCCCGGCCGAGCATGAGGTGCA	0.9504	1630	56

Continued on Next Page. . .

miRNA	miRNA*	p-value	miRNA abundance	miRNA* abundance
TGAGATGGTAATAACGGTGAT	CATCACCGTTATTACCACCTGG	0.9492	85	8
TGATTGAGCCGTGCCAATAAAA	TATTGGTGCGGTTCAATGAG	0.9486	22	5
CGGCTCCCGGCAGACGCACCA	TACGTCTGCCTGGGCGTCACGC	0.9485	21	7
CGCTTGGTGCAGGTCGGGAC	TCCCGGCCGAGCATGAGGTGC	0.9479	84	9
CAGGTGCCACTTATTTATGA	CCATAAATAAGTGGCACATGTG	0.9469	23	13
TCTGGATGACACAGGTGCCACC	AATAAGTGGCACATGTGTCATC	0.9440	31	10
TCTGGATGACACAGGTGCCAC	AATAAGTGGCACATGTGTCATC	0.9436	332	10
TGAGATGGTAATAACGGTGATA	CATCACCGTTATTACCACCTGG	0.9425	37	8
TCGCTTGGTGCAGGTCGGGACA	TCCCGGCCGAGCATGAGGTGCA	0.9421	851	18
TCGGTGCAGATCTTGGTGGTA	TCCGCCCTAGGTGTGCACCGG	0.9419	50	15
TCCAGAAATTGTCGCCTTGGA	CCAAGGTGAACAGCCTCTGG	0.9418	20	6
TCTAAGTCAGAATCCGGGCTAG	TCTGGCTTGGAATTGGGCTTGG	0.9410	24	13
CGCTTGGTGCAGGTCGGGACC	TCCCGGCCGAGCATGAGGTGC	0.9409	41	27
TTGGTGCCCTCGAAGACTCTCG	GGGCCTTTGAGGTAGCACCAAC	0.9408	20	8
CAGGTGCCACTTATTTATGAA	CCATAAATAAGTGGCACATGTG	0.9403	517	10
ATTTACGTGCAAGCGCAGTTGA	ACTGCGCTTGCCCGTAAATGGA	0.9400	37	29
ATATTGGTGCGGTTCAATTAG	TGATTGAGCCGTGCCAATACC	0.9396	35	35
TGATTGAGCCGTGCCAATATAA	TGATGTTGGAATGGCTCAAT	0.9396	39	8
CGTCGCGGTGACCGCCTTGA	TTGGCTAAGTCGCCGCGACGG	0.9393	22	16
TCGCTTGGTGCAGGTCGGAC	TGCCGACCTGCAGTAGGGGCC	0.9392	27	9
TCGCTTGGTGCAGGTTGGGAC	CCCGCCTTGCATCAACTGAAT	0.9386	161	27
TTTCCAATTCCACCCATTCCTA	GGGATCGGTGAGTTGGAAAGC	0.9382	21	15
TCGCTTGGTGCAGGTCGGGACT	TCCCGGCCGAGCATGAGGTGCA	0.9373	451	56
ACTAGGCTGTGTCTGGATGC	TCATCCAGACACGTTGTCCTAA	0.9363	40	34
CGTTTCAAGACATGTTGCCTG	TAGGACAACATGTCTGGATGA	0.9338	38	16
TGTTGGGCCCAAGCCTGTTAG	TAAAAGATTGGGCCGAATAA	0.9337	35	6
GCGCTCCGGACGCTGGCCTGTG	GCGTGGTGTCCGGTGCGCTC	0.9337	20	14
TCGTGACCTGCATGGGCCACCA	TGGCTTGGTGCAGGTCGGGAC	0.9320	56	6
TCGGGACCTGCATGGGCCACCA	TGGTGTCCGGTGCGCTCCCGG	0.9316	26	14
CAGGTGCCACTTATTTATGAAA	CCATAAATAAGTGGCACATGTG	0.9304	429	10
CGTGACCTGCATGGGCCACCA	TGGCTTGGTGCAGGTCGGGAC	0.9300	135	14
TTTACCAGGCACCCAGCAATG	CCATGGCTGGGAGCTGCCTGA	0.9296	242	16
CGGGACCTGCATGGGCCACCA	TGGTGTCCGGTGCGCTCCCA	0.9216	70	14
CAGGTGCCACTTATTTATGAAT	CCATAAATAAGTGGCACATGTG	0.9215	38	10
TTTCTCTGGTGCTTACTCAAC	TGAATGAAGTACTCAGAGAA	0.9214	20	20
CTGGATGACACAGGTGCCAC	AATAAGTGGCACATGTGTCATC	0.9195	32	10
GGGTTACGGTGCCAAACTGCG	TTGGCATGGTAGCCCTATAA	0.9191	43	15
GCTCCGGACGCTGGCCTGTGG	CACGTCGCGTGGTGTCCGGA	0.9184	22	19
TGAGATGGTAATAACGGTGA	CATCACCGTTATTACCACCTGG	0.9171	47	8
TGCGCCCGCCGTCCGCTTGCC	GAGCGGCCGTCGGTGCAGAT	0.9146	29	8
CTTCAAAGTTTCCGACGGGTGC	CACTTTGTTGGTGACTTTGAA	0.9138	28	5
TGTGGGTGGGGTGGAAAGATT	TCCATCCCATGTACATCAAT	0.9135	33	9
CGTGGGGGCATGTGATTGAA	TCAACCATGTGCCACTCTCGA	0.9113	296	36
GTTACGGTGCCAAACTGCGC	TCAGTTTGGTGTTGCTGGCAA	0.9093	29	12
TTCAACCTAGTACGAGAGGAA	CTCTCGTTTACGGTTGAGTCA	0.9083	24	7
TTCGTGCCTCATTGTGATGAT	GAAATCATCAAAAATGAGGCA	0.9053	272	7
TTGCCTTGGTGCCCTCGAAGAC	TTTTCAAGGGCCGCCGGGAGCA	0.9049	121	12
TCGCTTGGTGCATGTCGGGAC	TGCCTCATGTACCAACGAAAG	0.8998	49	7
TTGCTTGGTGCAGGTCGGGAC	TTTCCCGGCTGGTGCACCAA	0.8972	136	7
TCGCTTGGCGCAGGTCGGGAC	TTCTCGTGGTGCGTCAAGCAG	0.8962	35	14

Appendix E

E.1 *SBM* Results

The following tabes show the full results of comparison between *SBM* and `miRanda` from Chapter 7.

Target	LOO score	$\geqslant$ LOO score	miRanda(s)	$\geqslant$ miRanda(s)	miRanda(e)	$\geqslant$ miRanda(e)	$\geqslant$ miRanda(se)
ZK792.6/247-264	0.9587762	3561	109	83402	-14.10	38347	19155
F38A6.1a/271-288	1.0000000	1708	92	390274	-10.15	273523	159046
C18D1.1.1/526-542	0.9064860	10458	123	25838	-11.21	170342	19261
ZK792.6/666-683	0.9588374	3522	120	37964	-14.88	24456	8034
ZK792.6/458-475	0.9286717	7311	127	15875	-13.18	63901	9280
F38A6.1a/133-150	0.8742177	19177	108	95809	-14.56	29677	16710
C01G8.9a/21-38	0.8496634	23906	123	25838	-15.23	19926	5901
ZK792.6/132-148	0.8591035	20570	112	66769	-12.00	117457	33464
C01G8.9a/159-175	0.8134508	30895	113	61153	-13.33	58770	21054
ZK792.6/190-207	0.8068349	41812	91	403453	-13.33	58770	46851
C12C8.3a/693-709	0.7908144	39369	151	888	-25.13	14	9
C12C8.3a/742-757	1.0000000	1499	163	84	-25.06	16	4
ZK792.6/484-499	0.8981500	10232	107	102609	-17.86	3488	2822
F11A1.3a/1007-1021	0.9483082	4658	128	13871	-17.91	3345	1340
ZK792.6/343-361	0.9552215	4352	113	61153	-13.94	41961	16950
Mean	0.9032357	14869	119	92332	-15.46	60266	23992

Table E.1: Summary of the results for let-7 target predictions in *Caenorhabditis elegans*. Column “target” lists accession of the location of the validated target (UTR accession, start and end position), column “LOO score” shows the score for that target when left out of the *SBM*, column “$\geqslant$ LOO score” shows the number of regions scoring equal to or greater than the left out sequence, column “miRanda(s)” shows the raw score of the miRanda hit corresponding to the target region, column “$\geqslant$ miRanda(s)” shows the number of regions scoring equal to or greater than the target region, column “miRanda(e)” shows the minimum free energy (MFE) of the miRanda hit corresponding to the target region, column “$\geqslant$ miRanda(e)” shows the number of regions with an equal or more stable MFE than the target region, column “$\geqslant$ miRanda(se)” shows the number of regions with an equal or more stable MFE and a score greater than or equal to the target region.

Target	LOO score	$\geqslant$ LOO score	miRanda(s)	$\geqslant$ miRanda(s)	miRanda(e)	$\geqslant$ miRanda(e)	$\geqslant$ miRanda(se)
ZK792.6/126-148	0.8042310	4970	103	149244	-7.50	432158	114090
ZK792.6/187-207	0.5515072	132626	84	612304	-10.95	74633	61242
ZK792.6/249-264	0.9471851	355	114	55099	-10.61	91546	23993
ZK792.6/342-361	0.7607794	12552	102	160281	-8.80	235444	88454
ZK792.6/460-475	0.8575555	2012	126	15024	-10.20	114702	12066
ZK792.6/479-499	0.7394901	18375	94	317848	-12.08	38304	27872
ZK792.6/665-683	0.7264934	15846	122	25051	-11.22	64174	12052
Mean	0.7696059	26677	106	190693	-10.19	150137	48538

Table E.2: Summary of the results for miR-84 target predictions in *Caenorhabditis elegans*. Column “target” lists accession of the location of the validated target (UTR accession, start and end position), column “LOO score” shows the score for that target when left out of the *SBM*, column “$\geqslant$ LOO score” shows the number of regions scoring equal to or greater than the left out sequence, column “miRanda(s)” shows the raw score of the miRanda hit corresponding to the target region, column “$\geqslant$ miRanda(s)” shows the number of regions scoring equal to or greater than the target region, column “miRanda(e)” shows the minimum free energy (MFE) of the miRanda hit corresponding to the target region, column “$\geqslant$ miRanda(e)” shows the number of regions with an equal or more stable MFE than the target region, column “$\geqslant$ miRanda(se)” shows the number of regions with an equal or more stable MFE and a score greater than or equal to the target region.

Target	LOO score	$\geqslant$ LOO score	miRanda(s)	$\geqslant$ miRanda(s)	miRanda(e)	$\geqslant$ miRanda(e)	$\geqslant$ miRanda(se)
CG12487.3/223-241	0.9464120	94	164	215	-22.76	127	44
CG5185.3/279-297	1.0000000	34	173	36	-24.25	45	10
CG3096.3/152-170	1.0000000	34	168	117	-24.14	46	16
CG12487.3/250-268	1.0000000	34	179	8	-24.71	38	6
CG3166.3/1100-1118	0.9505960	76	140	3490	-18.77	2322	621
CG6096.3/103-121	1.0000000	34	172	47	-23.80	63	12
CG8346.3/78-96	0.9659194	58	185	2	-28.03	2	1
CG5185.3/334-352	1.0000000	34	170	83	-23.24	93	25
CG6494.3/447-465	0.9188494	155	179	8	-25.24	25	3
CG6096.3/24-42	1.0000000	34	171	64	-23.71	72	19
CG6096.3/68-86	0.9614793	65	170	83	-23.71	72	21
CG8328.3/63-81	0.7726001	2015	145	2210	-16.48	10689	1001
CG3166.3/1586-1602	0.8547854	393	138	4162	-16.44	10990	1630
CG3166.3/29-46	0.8454907	513	130	11501	-16.84	8407	2226
CG3166.3/1294-1312	0.8607346	521	108	111001	-13.26	75409	26300
Mean	0.9384578	273	159	8868	-21.69	7227	2129

Table E.3: Summary of the results for mir-7 target predictions in *Drosophila melanogaster*. Column “target” lists accession of the location of the validated target (UTR accession, start and end position), column “LOO score” shows the score for that target when left out of the *SBM*, column “$\geqslant$ LOO score” shows the number of regions scoring equal to or greater than the left out sequence, column “miRanda(s)” shows the raw score of the miRanda hit corresponding to the target region, column “$\geqslant$ miRanda(s)” shows the number of regions scoring equal to or greater than the target region, column “miRanda(e)” shows the minimum free energy (MFE) of the miRanda hit corresponding to the target region, column “$\geqslant$ miRanda(e)” shows the number of regions with an equal or more stable MFE than the target region, column “$\geqslant$ miRanda(se)” shows the number of regions with an equal or more stable MFE and a score greater than or equal to the target region.

Target	LOO score	$\geqslant$ LOO score	miRanda(s)	$\geqslant$ miRanda(s)	miRanda(e)	$\geqslant$ miRanda(e)	$\geqslant$ miRanda(se)
CG6096.3/135-154	0.7550454	3118	143	735	-11.44	26630	491
CG8328.3/27-45	1.0000000	8	135	3154	-8.69	128342	2427
CG3096.3/33-52	0.9287263	161	130	6986	-7.00	305330	5993
CG3096.3/138-157	0.8767518	473	136	2730	-7.30	264578	2638
CG5185.3/46-65	0.9598055	64	139	1586	-8.84	118531	1390
CG12487.3/188-208	0.8200697	1298	112	56515	-8.54	138880	18111
CG12487.3/62-82	0.8714641	627	127	10317	-10.03	61435	3372
CG6096.3/210-230	0.9076755	207	128	9883	-6.26	429349	8181
Mean	0.8899422	745	131	11488	-8.51	184134	5325

Table E.4: Summary of the results for mir-4 target predictions in *Drosophila melanogaster*. Column “target” lists accession of the location of the validated target (UTR accession, start and end position), column “LOO score” shows the score for that target when left out of the *SBM*, column “$\geqslant$ LOO score” shows the number of regions scoring equal to or greater than the left out sequence, column “miRanda(s)” shows the raw score of the miRanda hit corresponding to the target region, column “$\geqslant$ miRanda(s)” shows the number of regions scoring equal to or greater than the target region, column “miRanda(e)” shows the minimum free energy (MFE) of the miRanda hit corresponding to the target region, column “$\geqslant$ miRanda(e)” shows the number of regions with an equal or more stable MFE than the target region, column “$\geqslant$ miRanda(se)” shows the number of regions with an equal or more stable MFE and a score greater than or equal to the target region.