UNIVERSITA' DEGLI STUDI DI MILANO

Dipartimento di Scienze Biomolecolari e Biotecnologie

Scuola di dottorato in
Scienze Biologiche e Molecolari

Corso di dottorato in
Biologia Cellulare e Molecolare XXIII CICLO

TESI DI DOTTORATO DI RICERCA

MicroRNA Discovery and Characterization in *Vitis vinifera* Using smallRNA Deep Sequencing and Support Vector Machines

VIVIANA PICCOLO

R07589 BIO/11

Tutor CARMELA GISSI

Co-tutor DAVID S. HORNER

Coordinatrice Chiar.ma Prof.ssa GIULIANA ZANETTI

Anno Accademico 2009-2010

This thesis is dedicated to my family

Abstract

Introduction

MiRNAs are small non coding RNAs that play an important role in the regulation of multiple cell events. They inhibit gene expression at post transcriptional level by binding mRNA targets that are degraded or squestred from translation.

Vitis vinifera is the first whole genome sequenced for a commercially important fruit species. Here we present the development and implementation of diverse strategies for the identification and validation of miRNAs of the grapevine. Many putative conserved microRNA precursors were identified by comparative methods and subsequently validated through high throughput smallRNA sequencing and oligonucleotide array technology. Additional bioinformatics tools were implemented for the *ab-initio* prediction of miRNAs and for the identification of *lineage-specific* miRNAs from smallRNA deep sequence data.

Materials and methods

Software to assist in the design of oligonucleotide arrays for the validation of miRNA expression in grape was developed and oligonucleotide array and deep sequencing experiments were used to confirm the expression of conserved mature miRNAs from most of these loci in at least one tissue or developmental stage.

Support Vector Machine - based software to predict *novel* miRNAs and to study their evolution was developed and shown to outperform similar published methods. This classifier was also incorporated into a novel approach to the analysis of smallRNA deep sequence utilizing patterns of mapping of reads on the genome. Our method performs well in the identification novel miRNAs and non-canonical miRNA-like loci.

Results

Many conserved miRNAs were identified and show strong patterns of tissue specific expression. We have shown that for many, but by no means all known miRNA precursors, evidence for primary transcript expression can be obtained from high throughput transc-riptome analysis, classically performed to follow expression levels of protein coding genes. We estimated patterns of splicing and alternative splicing of known pri-miRNA transcripts

The method developed for the identification of plant miRNA precursors from smallRNA NGS data recovers many novel, canonical miRNAs from *Vitis* and is capable of identifying

loci producing miRNA-like smallRNAs with characteristics that are atypical of most conserved miRNAs.

The patterns of smallRNA generated from putatively *lineage specific* loci have been considered in the context of a current model of miRNA gene evolution.

Index

Preface

This thesis contains work performed as part of a large collaborative effort aimed principally at the characterization of the genome of the grapevine, *Vitis vinifera* L., which was sequenced, assembled and annotated by the French Italian Public Consortium for Grapevine Genome Characterization (Jaillon, Aury et al. 2007).
In particular, the work presented here focuses on the implementation of different bioinformatics strategies for the identification of both conserved and novel miRNAs in this species.

The first chapter introduces the topics of the thesis, while the following chapters describe the results obtained during these three years of my Ph.D.

The first results chapter describes comparative prediction of conserved miRNAs in *Vitis* and the use of high throughput methods to validate their expression. This work was pubblished in BMC Genomics (Mica, Piccolo et al. 2010).

The following chapter concerns the implementation of an *ab-initio* pre-miRNA prediction tool, while the final chapter outlines novel strategies for the interpretation of smallRNA deep sequence data and their application to the discovery of miRNAs in the grapevine genome.

Chapter 1

1.1 Discovery of microRNAs

In 1993, it was discovered that *lin-4*, a gene known to control the timing of *Caenorhabditis elegans* larval development, did not code for a protein, but, instead, generated two small RNAs of different size (one of 22 nt and the other of about 61 nt) (Lee, Feinbaum et al. 1993). The longer RNA was predicted to fold into a stem loop structure and was proposed to be the precursor of the shorter one. These *lin-4* RNAs had antisense complementarity to multiple sites in the 3'UTR of the *lin-14* gene (Lee, Feinbaum et al. 1993; Wightman, Ha et al. 1993) . A short region in the 3' UTR of *lin-14* was required for the repression of *lin-14* by the *lin-4* gene product (Lee, Feinbaum et al. 1993). In 1993 was noted a reduction of the amount of LIN-14 protein *without* noticeable change in levels of *lin-14* mRNA. This observation has created the model of action whereby substantially lin-*4* RNAs (non gene product) pairs to the *lin-14* 3'UTR to specify the translational repression of the *lin-14* message. This negative regulation triggers the transition from cell divisions of the first larval stage to those of the second (Lee, Feinbaum et al. 1993; Wightman, Ha et al. 1993). In *C. elegans*, cell lineages have distinct characteristics during 4 different larval stages (L1–L4) (Fig.1.1) and mutations in *lin-4* disrupt the temporal regulation of larval development, causing L1 (the first larval stage)-specific cell-division patterns to reiterate at later developmental stages. Instead, in worms deficient for *lin-14* was oserved the opposite developmental phenotypes (the omission of the L1 cell fates and the premature development into the L2 stage).

Subsequently another non-coding RNA was discovered: *let-7* RNA, that is involved in the regulation of larval development. *let-7* RNA promotes the transition from late-larval to adult cell fates in the same way that the *lin-4* RNA acts to activates the progression from the first larval stage to the second (Reinhart, Slack et al. 2000; Slack, Basson et al. 2000). Furthermore homologs of the *let-7* gene were soon identified in the human and fly genomes, and *let-7* RNA itself was detected in human, *Drosophila*, and eleven other bilateral animals (Pasquinelli, Reinhart et al. 2000).

Because of their common roles in controlling the timing of developmental transitions, initially the *lin-4* and *let-7* RNAs were called *small temporal RNAs* (stRNAs) (Pasquinelli, Reinhart et al. 2000) and only later was identified as members of new class of tiny (20-25 nt)

regulatory RNAs (Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001; Lee and Ambros 2001). The term microRNA was subsequently used to refer to these stRNAs and to all the other tiny RNAs with similar features but unknown functions (Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001; Lee and Ambros 2001). Small RNA cloning efforts from flies, worms, human, plants cells revealed numerous additional miRNAs (Lagos-Quintana, Rauhut et al. 2001; Mourelatos, Dostie et al. 2002; Ambros 2003; Aravin, Lagos-Quintana et al. 2003; Dostie, Mourelatos et al. 2003; Houbaviy, Murray et al. 2003; Lim, Lau et al. 2003).

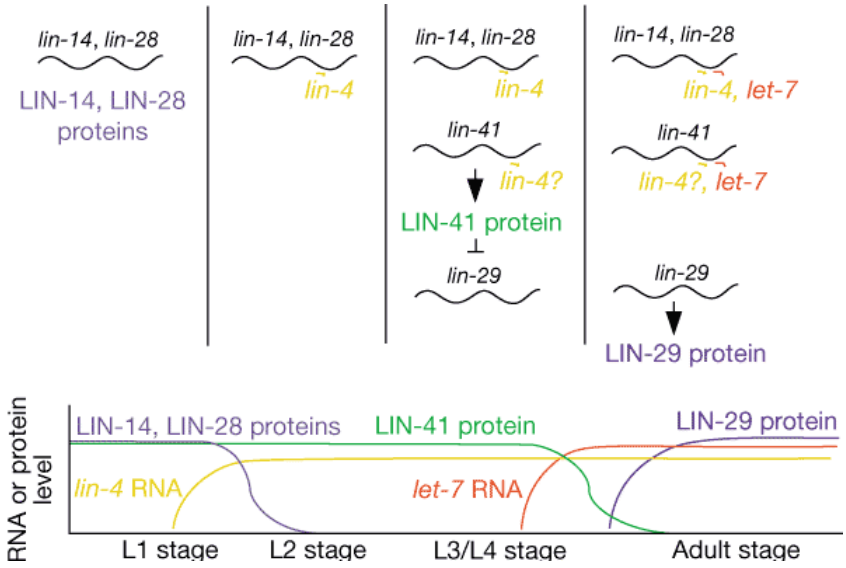


Fig.1.1 - A model of successive regulation of heterochronic gene activities by *lin-4* and *let-7* RNAs
LIN-14 and LIN-28 expression levels are reduced by *lin-4* RNA expression at end of 1st larval stage, allowing progression to late larval stages.
In late larval stages, expression of LIN-41 and other genes may be similarly downregulated by *let-7* RNA, relieving their repression of LIN-29 protein expression, allowing progression to adult stage.
Because the *lin-29* mRNA does not contain sites complementary to the *let-7* RNA, *lin-29* is not likely to be a direct target of *let-7*.

The spread and importance of miRNA-directed gene regulation are coming into focus as more miRNAs and their regulatory targets and functions are discovered. MiRNA functions include control of cell proliferation, cell death, and fat metabolism in flies (Brennecke and Cohen 2003; Xu, Vernooy et al. 2003) neuronal patterning in nematodes (Johnston and Hobert 2003) modulation of hematopoietic lineage differentiation in mammals (Chen, Li et al. 2004) and control of leaf and flower development in plants (Aukerman and Sakai 2003; Emery, Floyd et al. 2003; Palatnik, Allen et al. 2003; Chen, Li et al. 2004)

A registry has been set up to catalog the miRNAs and facilitate the naming of newly identified miRNA genes (Griffiths-Jones 2006). Apart from in animals, miRNAs have been identified in plants, algae and many protist species, although to date there is no evidence for their occurrence in fungi. At the time of writing, over 1000 (1048) miRNAs have been described in humans, 176 in *Drosophila melanogaster*, 175 in *Caenorhabditis elegans*, 213 in *Arabidopsis thaliana*, 462 in *Oryza sativa* and 229 in the basal land plant *Physcomitrella patens* (as well as numerous others in other species). It is thus clear that miRNAs are a diffuse and important gene regulatory mechanism in eukaryotic organisms.

1.2 General introduction to microRNAs

MicroRNAs (miRNAs) are small single-stranded RNAs (~19-25 nt) generated from endogenous transcripts that can form local hairpin structures named miRNA precursors (Ambros 2003).

The majority of miRNA genes are located in intergenic regions or in antisense orientation to annotated genes (Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001; Lee and Ambros 2001; Mourelatos, Dostie et al. 2002) (Fig.1.2).

The fact that many miRNA genes come from regions of the genome quite distant from previously annotated genes, implies that they derive from independent transcription units (Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001; Lee and Ambros 2001).

Nonetheless, a sizable minority, particularly in animals, are in the introns of pre-mRNAs (Lin, Miller et al. 2006; Li, Tang et al. 2007; Ruby, Jan et al. 2007; Barik 2008; Golan, Levy et al. 2010; Hsu, Lin et al. 2010). These are preferentially in the same orientation as the predicted mRNAs, suggesting that most of these miRNAs are not transcribed from their own promoters but are instead processed from the introns, as seen also for many snoRNAs (Aravin, Lagos-Quintana et al. 2003; Lagos-Quintana, Rauhut et al. 2003; Lai, Tomancak et al. 2003; Lim, Glasner et al. 2003; Li, Tang et al. 2007).

This arrangement provides a convenient mechanism for the coordinated expression of a miRNA and a protein. Regulatory scenarios are easy to imagine in which such coordinate expression could be useful, which would explain the conserved relationships between miRNAs and host mRNAs. A striking example of this conservation involves *mir-7*, found in the intron of *hnRNP K* in both insects and mammals (Aravin, Lagos-Quintana et al. 2003).

Other miRNA genes are clustered in the genome with an arrangement and expression pattern implying transcription as a multi-cistronic primary transcript (Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001) (Fig.1.2 and Fig.1.3)

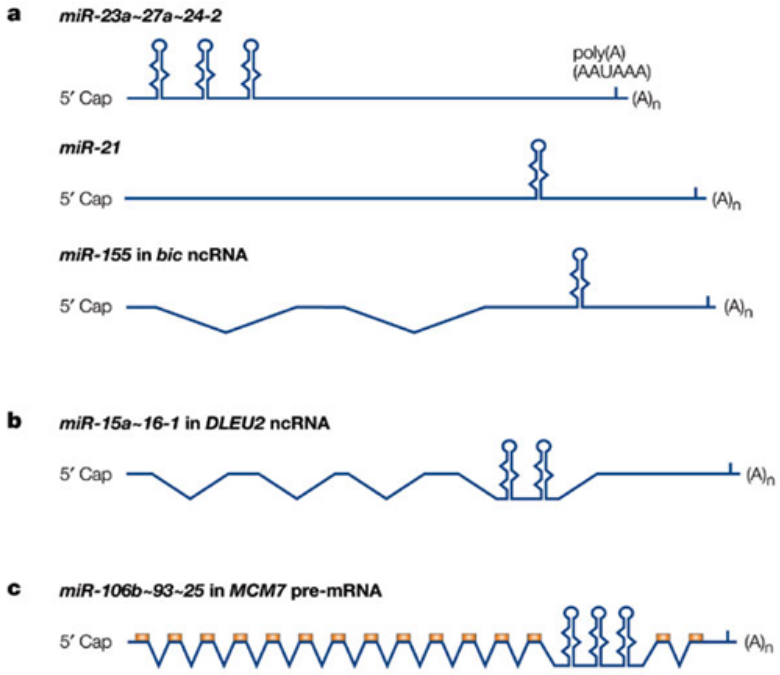


Fig.1.2 - MiRNA genomic position
miRNAs can be categorized into three groups according to their genomic locations relative to their positions in an exon or intron.
(a) Exonic miRNAs in non-coding transcripts such as an miR-23a~27a 24-2 cluster, miR-21 and miR-155. miR-155 was found in a previously defined non-coding RNA (ncRNA) gene, bic (Tam 2001).
(b) Intronic miRNAs in non-coding transcripts. For example, an miR-15a~16-1 cluster was found in the fourth intron of a previously defined non-coding RNA gene, DLEU2 (Calin, Dumitru et al. 2002)
(b) Intronic miRNAs in protein-coding transcripts. For example, an miR-106b~93~25 cluster is embedded in the thirteenth intron of DNA replication licensing factor MCM7 transcript (variant 1, which encodes isoform 1). The mouse miR-106b~93~25 homologue is also found in the thirteenth intron of the mouse MCM7 homologue gene. (Rodriguez, Griffiths-Jones et al. 2004) The hairpins indicate the miRNA stem-loops. Orange boxes indicate the protein-coding region. This figure is not to scale. From (Kim 2005)

Although the majority of worm and human miRNA genes are isolated and not clustered (Lim, Glasner et al. 2003; Lim, Lau et al. 2003) over half of the known *Drosophila* miRNAs are clustered (Aravin, Lagos-Quintana et al. 2003). MiRNAs within a genomic cluster are often, though not always, related to each other; and related miRNAs are sometimes, but not always, clustered (Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001). Orthologs of *C. elegans lin-4* and *let-7* are clustered in the fly and human genomes and are coexpressed, sometimes from the same primary transcript, leading to the idea that

the genomic separation of *lin-4* from *let-7* in nematodes might be unique to the worm lineage (for *let-7* see Fig.1.3B) (Aravin, Lagos-Quintana et al. 2003; Bashirullah, Pasquinelli et al. 2003; Sempere, Sokol et al. 2003).

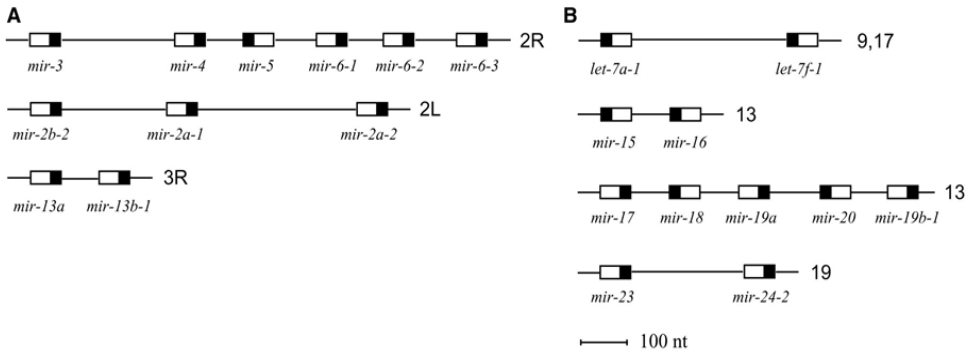


Fig.1. 3 - Genomic organization of miRNA gene clusters.
The precursor structure is indicated as a box, and the location of the miRNA within the precursor is shown in black; the chromosomal location is also indicated to the right.
(**A**) *D. melanogaster* miRNA gene clusters.
(**B**) Human miRNA gene clusters. The cluster of *let-7a-1* and *let-7f-1* is separated by 26,500 nt from a copy of *let-7d* on chromosomes 9 and 17. A cluster of *let-7a-3* and*let-7b*, separated by 938 nt on chromosome 22, is not illustrated. From Lagos-Quintana, 2001 modified

1.3 Other classes of smallRNAs.

The intense scientific interest in microRNAs quickly lead to the realization that miRNAs constitute only a fraction of the physiologically relevant smallRNAs produced in animal and plant cells. In animals, other classes of smallRNAs, such as endogenous-siRNA and PIWI-associated siRNAs are involved in the suppression of transposon mobility (Kim, Han et al. 2009) while smallRNAs are often produced specifically from snoRNAs (Bachellerie, Cavaille et al. 2002; Ender, Krek et al. 2008; Saraiya and Wang 2008; Taft, Glazov et al. 2009) and tRNAs (Taft, Glazov et al. 2009; Brameier, Herwig et al. 2010). In addition, smallRNAs are produced from promoter regions of actively transcribed genes (Taft, Kaplan et al. 2009). The function and biogenetic mechanisms of these molecules remain poorly understood and other classes of smallRNA will likely be characterized in the future. However, it should be noted that a large fraction of small RNAs are not microRNAs, a consideration that complicated the annotation of real miRNAs.

In plants, the situation seems to be even more complicated. The majority of small RNA molecules observed to date are derived from transposons and other repetitive elements in the genome. These smallRNAs do not derive from hairpin precursors, but from transcripts which

have been converted to dsRNA by RNA-dependent RNA polymerases. These smallRNAs are thought to direct genome methylation in *cis* and to repress transcription of repetitive elements (Xie and Qi 2008). SmallRNAs (nat-siRNAs) can also be derived from complementary antisense transcripts (Borsani, Zhu et al. 2005; Wang, Chua et al. 2006), while promoter derived and snoRNA derived smallRNAs are also observed. Trans-Acting siRNAs (ta-siRNAs) are produced through a complicated mechanism involving miRNA targeting of specialized non-coding primary transcripts and RNA-dependent RNA polymerase activities (Allen, Xie et al. 2005; Allen and Howell 2010). These small RNAs behave rather like miRNAs and go on to target other mRNAs. Additionally, in all systems studied to date, a significant proportion of the smallRNA molecules characterized by sequencing-based strategies are derived from degradation of ribosomal RNA and even of mRNA molecules. Recent studies also suggest that significant overlap between biogenetic pathways of smallRNAs can occur, meaning that some loci produce molecules with characteristics of more than one class of smallRNA (Vazquez, Legrand et al. 2010).

1.4 Biogenesis of miRNAs

1.4.1 Transcription

Our knowledge of miRNA biogenesis has significantly advanced in recent years. However, little is known about transcription of miRNA genes although it is likely to be the key regulatory step in miRNA biogenesis.

The biogenesis of miRNAs begins with the transcription of a primary transcript (pri-miRNA), an hairpin which can be up to several kilobases in length (Lee, Kim et al. 2004) (Fig.1.4).

MiRNAs genes might either be transcribed by RNA pol-II or pol-III. In general, Pol II produces the mRNAs and some noncoding RNAs, including the small nuclear RNAs (snoRNAs) and four of the small nuclear RNAs (snRNAs) of the spliceosome, whereas Pol III produces some of the shorter noncoding RNAs, including tRNAs, 5S ribosomal RNA, and the U6 snRNA. The miRNAs processed from the introns of protein-coding host genes are undoubtedly transcribed by Pol II (Zeng, Wagner et al. 2002; Zeng and Cullen 2003; Borchert, Lanier et al. 2006).

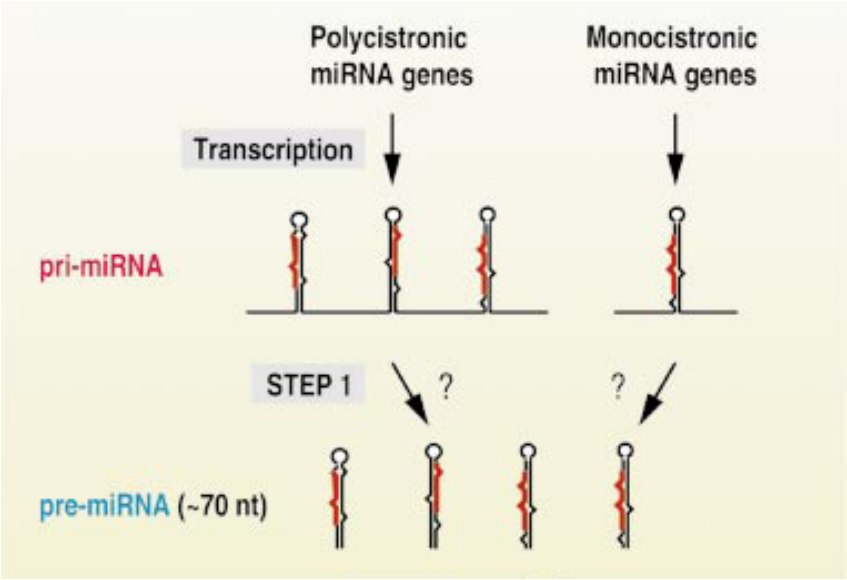


Fig.1.4 - A model for miRNA biogenesis.
MiRNA genes are transcribed by an unidentified polymerase to generate the primary transcripts, referred to as pri-miRNAs. Illustrated in the upper left is the clustered miRNA, such as miR-23~27~24-2, of which the pri-miRNA is polycistronic. Illustrated in the upper right is the miRNA, such as miR-30a, of which the pri-miRNA is monocistronic. The first-step processing (STEP 1) by a RNA pol-II or pol-III results in pre-miRNAs of ~70 nt From (Lee, Jeon et al. 2002) modified

The following observations provide indirect evidence that many of the other miRNAs are pol II products, even though most of metazoan miRNA genes do not have the classical signals for polyadenylation (Ohler, Yekta et al. 2004):

(1) the pri-miRNAs can be quite long, more than one 1 kb, which is longer than typical pol III transcripts;

(2) these presumed pri-miRNAs often have internal runs of uridine residues, which would be expected to prematurely terminate pol III transcription;

(3) many miRNAs are differentially expressed during development, as is observed often for pol II but not pol III products;

(4) fusions that place the open reading frame of a reporter protein downstream from the 5' portion of miRNA genes lead to robust reporter protein expression, suggesting that miRNA primary transcripts are capped pol II transcripts. Examples of such fusions include artificial reporter constructs designed to investigate the regulation of miRNA expression (Johnson, Lin et al. 2003; Johnston and Hobert 2003) and a natural chromosome translocation linked to an aggressive B cell leukemia, in which a truncated *MYC* gene is fused to the 5' portion of miR-142 (Gauwerky, Huebner et al. 1989; Lagos-Quintana, Rauhut et al. 2002).

Although these observations indicate that many miRNAs are pol II transcripts, others might still be pol III transcripts, just as most but not all snRNAs are pol II products.

Ectopic expression of miR-142 and other miRNAs from a pol III promoter produces efficiently and precisely processed miRNAs that function in *vivo* (Chen, Li et al. 2004), indicating that there is no obligate link between the identity of the polymerase and downstream miRNA processing or function.

1.4.2 Pri-miRNA processing

Pri-miRNA processing is a critical step in miRNA biogenesis. This initial processing event predetermines mature miRNA sequences by generating one end of mature miRNA (Lee, Ahn et al. 2003; Lund, Guttinger et al. 2004) and releasing the characteristic hairpin precursor (pre-miRNA) from the primary transcript.

Both in pants and in animals pri-miRNA processing is mediated by RNase III type enzymes. In animals the protein is called Drosha, while in plants DCL1.

In general, RNase III proteins are grouped into three classes based on their domain organization:

(a) class I proteins include RNase III proteins that are present in bacteria and yeasts. Each proteins contains one RNase III domain (RIIID) and one double stranded RNA (dsRNA)-binding domain (dsRBD);

(b) class II proteins such as Drosha possess two RIIIDs and a dsRBD. Drosha homologs are present only in animals. These proteins are large (130–160 kDa) and possess extended N-termini whose functions are unknown. The N-terminal portion of Drosha contains a proline-rich region as well as a serine/arginine-rich region;

(c) class III (Fig.1.5) includes Dicer homologs that are conserved in *Schizosaccharomyces pombe*, plants, and animals. DICER homologs are more or less 200 kDa and contain multiple domains. Apart from two RIIIDs and a dsRBD, DICER has a long N-terminus containing a DExH RNA helicase/ATPase domain, as well as DUF283 and the PAZ domain. The PAZ domain is also found in a group of highly conserved proteins, referred to as Argonaute proteins (also known as PPD proteins). Structural and biochemical studies of the PAZ domain from *Drosophila* Ago1 and Ago2 suggest that the PAZ domain binds to the 3' protruding end of small RNAs (Lingel, Simon et al. 2003; Song, Liu et al. 2003; Yan, Yan et al. 2003). The function of the other domains in DICER are not yet clear. Although DICER associates with several other proteins (Argonaute proteins in various organisms, RDE-4 in *Caenorhabditis elegans*, R2D2 in *Drosophila*, and dFMR1 in *Drosophila*) (Hammond, Boettcher et al. 2001; Ishizuka, Siomi et al. 2002; Tabara, Yigit et al. 2002), these interacting

proteins do not seem to be required for the cleavage reaction itself because purified human DICER and *Drosophila* DICER-2 can catalyze the cleavage reaction .

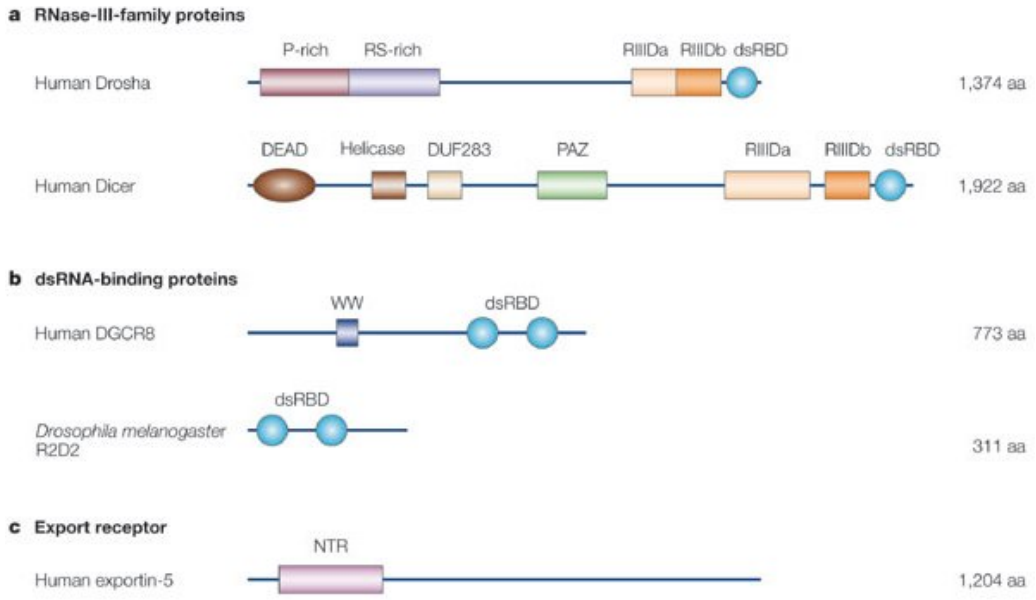


Fig.1.5 - RNase III type enzymes domains
(a) The RNase III domain (RIIID) is the catalytic domain that is responsible for the endonucleolytic reaction of RNase III enzymes such as DICER and Drosha. The RIIIDs (shown as RIIIDa and RIIIDb) are well conserved motifs found in RNase-III-type proteins of eubacterial, archaeal and eukaryotic origin. The double-stranded RNA-binding domain (dsRBD) is also a well conserved motif in many double-stranded RNA (dsRNA)-binding proteins of diverse functions, including Drosha and DICER. The biological significance of the proline-rich (P-rich) region is unknown. The RS-rich region is abundant in arginine and serine residues. The function of this region is also unclear, although the C terminus of this region was shown to be important for the activity of Drosha (Han, Lee et al. 2004). The PAZ domain binds to the 3' end of small RNAs (Lingel, Simon et al. 2004; Ma, Ye et al. 2004). The DEAD-box RNA helicase domain is typical of enzymes that hydrolyse ATP and unwind an RNA duplex. The DUF283 domain has no known function.
(b) The WW motif is known as a protein interaction module that binds to the P-rich domain, although the role of this domain in DGCR8 remains unclear.
(c) The nuclear transport receptor (NTR) domain is found in many Ran-dependent nuclear transport factors aa, amino acids (Nakielny and Dreyfuss 1999). From (Kim 2005) modified

1.4.2.1 *Pri-miRNA processing in animals*

Human Drosha fractionates more or less at 650 kDa, indicating that Drosha functions as part of a large complex. This complex is called Microprocessor and cuts the stem loop containing the future miRNA out of the pri-miRNA (Han, Lee et al. 2004) (Fig.1.6 and Fig1.7). In the Microprocessor complex, Drosha interacts with DGCR8 (Fig.1.6), that is a protein of unknown function which contains two dsRBDs and a putative WWdomain (Lee, Ahn et al. 2003; Han, Lee et al. 2004; Zeng, Yi et al. 2005).

As in the class III enzyme human DICER, the two RIIIDs of human Drosha form an intramolecular dimer where the two domains are distinct in their roles. The RIIIDa cuts the 3′ strand, while the RIIIDb cleaves the 5′ strand, independently of each other. This result suggests that the Drosha protein is capable of orienting itself on pri-miRNA in a way that each RIIID is positioned on the correct strand (Lee, Ahn et al. 2003; Han, Lee et al. 2004; Zeng, Yi et al. 2005; Han, Lee et al. 2006) (Fig1.6).

DGCR8 may help Drosha to be correctly positioned on pri-miRNA (Kim 2005). DGCR8 may provide such an RNA-binding module and thereby may serve as an essential component of the Drosha complex.

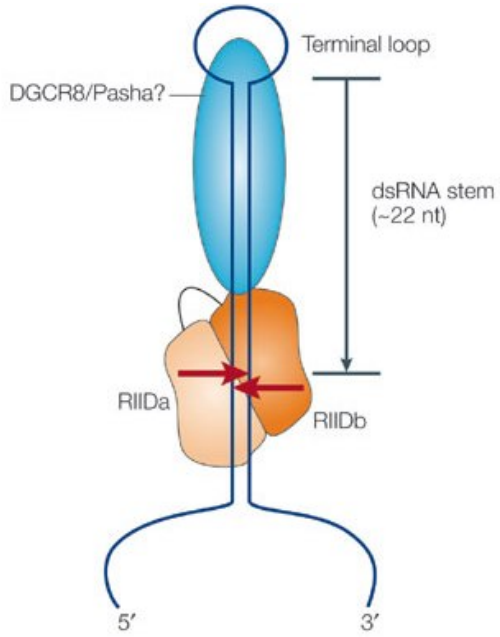


Fig1.6 - Pri-miRNA nuclear processing: Drosha.
The double-stranded RNA (dsRNA) substrate is laid on the cleft between RIIIDa (light orange) and RIIIDb (dark orange). The catalytic site on the RIIIDa side cleaves the 3' strand, whereas another catalytic site on the RIIIDb cleaves the 5' strand.
Drosha binds to a primary transcript (pri-miRNA) and introduces a cut at approximately two helical turns, ~22 nucleotides (nt), from the terminal loop. For simplicity, the domains of Drosha other than the RIIIDs are not shown. It was proposed that DGCR8/Pasha might help the binding of the complex to RNA and/or orienting the complex on pri-miRNA. From (Kim 2005) modified

The *DGCR8* gene was originally identified in the "DiGeorge syndrome chromosomal region (DGCR)" at human chromosome *22q11* (Shiohama, Sasaki et al. 2003). Monoallelic deletion of this region is associated with a complicated clinical phenotype, including DiGeorge syndrome/conotruncal anomaly face syndrome/velocardiofacial syndrome, although it remains unknown whether DGCR8 is involved in this genetic disorder.

Experiments suggest that Drosha and DGCR8 may be the only essential components of the pri-miRNA processing complex; however, the active complex is more or less 650 kDa, which may accommodate multiple subunits (Lee, Ahn et al. 2003; Han, Lee et al. 2004; Zeng, Yi et al. 2005; Han, Lee et al. 2006).

The precision of Drosha-DGCR8 cleavage is crucial for the fidelity of miRNA maturation: if the position of the Drosha cut is shifted by a single nucleotide on the pri-miRNA, then DICER cleavage, too, will be shifted, and the final miRNA will have different 5′ and 3′ ends (and potentially not be functional).

Initially, Drosha was thought to cut the stem by measuring two helical turns from the loop (Zeng et al., 2005), but other studies shown that the cut is more or less at position 11 bp from the basal segment and not from the terminal loop (Fig.1.7) (Han, Lee et al. 2004; Seitz and Zamore 2006).

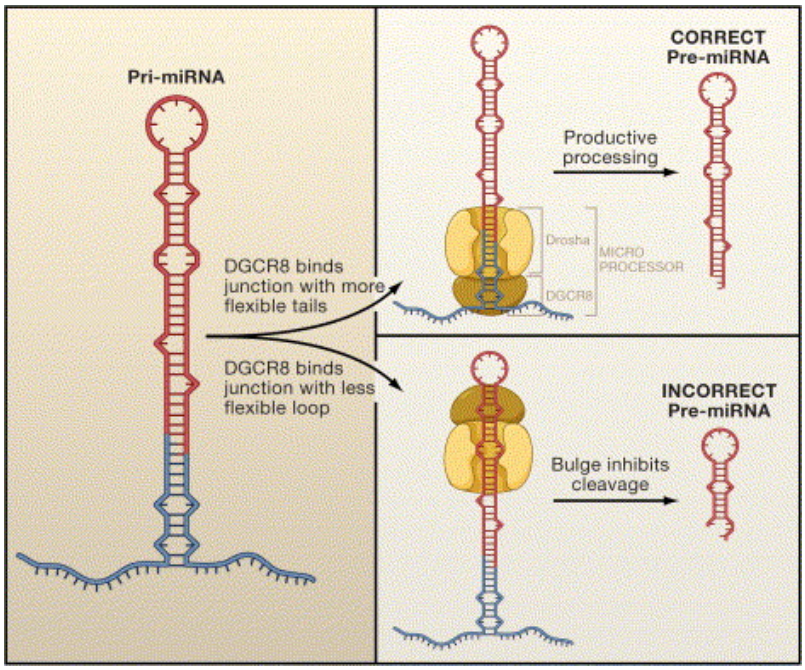


Fig. 1.7 - Pri-miRNA Processing by Drosha and DGCR8, Components of the Microprocessor
Structural features of pri-miRNAs promote their accurate processing into pre-miRNAs by the RNase III enzyme Drosha and its double-stranded RNA binding protein partner, DGCR8 (Pasha in invertebrates). DGCR8 is thought to bind more favorably to the junction between the rigid double-stranded stem and the 5′ and 3′ flexible, single-stranded segments of the pri-miRNA than to the junction between the stem and the considerably more constrained loop. Correct binding of DGCR8 to the base of the stem is proposed to position the processing center of Drosha ~11 bp up along the stem, where it makes a staggered pair of breaks in the RNA to create the ~65 nucleotide-long pre-miRNA. Binding of DGCR8 at the loop end of the stem positions Drosha inappropriately. Unpaired or weakly paired nucleotides at this site serve to discourage such unproductive leavage, reducing the number of abortive Drosha products and favoring accurate re-miRNA production. Taken from (Seitz and Zamore 2006) modified

The nature of the terminal loop seems not to be is particularly relevant to pri-miRNA processing, although both processing and DGCR8 binding were slightly impaired in a mutant with a small loop, suggesting that the presence of a large loop may be beneficial to some extent (Han, Lee et al. 2004). It is possible that a terminal loop that is too small in size may impose structural constrains upon the stem and affect processing (Han, Lee et al. 2004; Han, Lee et al. 2006).

Other studies demonstrated that the terminal loop can be replaced by single-stranded RNA with no major effect on pri-miRNA processing, but the single-stranded RNA segments flanking the base of the stem are indispensable for Drosha cleavage. In fact deleting these single-stranded regions or converting them to double-stranded RNA by annealing a synthetic oligonucleotide to them greatly impairs the conversion of pri-miRNA to pre-miRNA (Han, Lee et al. 2004; Zeng, Yi et al. 2005).

Modifying the length of the base of the stem also shifts the cleavage site. So it seems that the molecular ruler is anchored by the junction between the 5′ and 3′ single-stranded segments and the base of the double-stranded stem.

1.4.2.2 *Pri-miRNA processing in plants*

Processing of miRNA precursors in plants is not yet understood in the same level of detail as in animals, but cleavage of the primary transcript to the mature miRNA duplex is thought to be carried out by DCLs (DCER like proteins) (Chen 2005; Jones-Rhoades, Bartel et al. 2006).

In *Giardia intestinalis*, DICER (Fig.1.8) was shown to interact with 2-nt 3'-overhangs of dsRNAs through its PAZ domain (PIWI/AGO/ZWILLE), a highly-conserved RNA binding domain (RBD) also found in AGO proteins.

Both in animals and plants, the function of DICERs and DCLs on dsRNAs generally depends on interaction with DRB proteins which are thought to select the substrates (Jacobsen, Running et al. 1999; Golden, Schauer et al. 2002; Schauer, Jacobsen et al. 2002; Xie, Johansen et al. 2004) (Fig.1.9).

The positively charged flat helix which connects the PAZ domain to the catalytic center further stabilizes the dsRNA through electrostatic interactions.

The catalytic residues (*) of the two RNAse III domains (RIII) are shifted and cut one strand of the dsRNA each to release the 25-nt-long duplex with 2-nt overhangs (MacRae, Zhou et al. 2007).

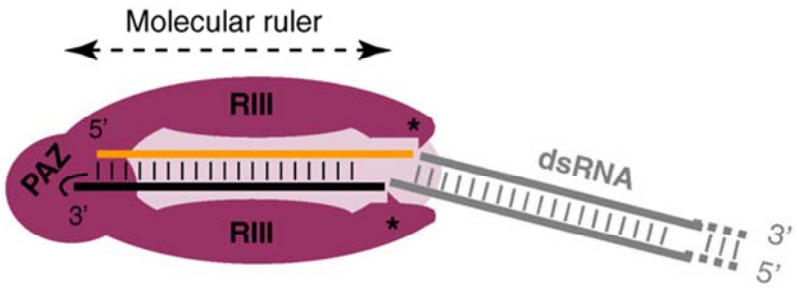


Fig.1.8 - Mechanism of action of DICER in *Giardia intestinalis*
The positively charged flat helix which connects the PAZ domain to the catalytic center further stabilizes the dsRNA through electrostatic interactions. The catalytic residues (*) of the two RNAse III domains (RIII) are shifted and cut one strand of the dsRNA each to release the 25-nt-long duplex with 2-nt overhangs (MacRae, Zhou et al. 2007). From (Vazquez, Legrand et al. 2010) modified

Plant DCLs probably function similarly. However, whether they generate an initial overhang for primo-interaction or whether they also interact with other types of free ends is unknown. Moreover, whether DRBs contribute to the primo-interaction remains to be determined. The physical distance between the PAZ domain and processing center forms a molecular ruler which determines the size of the sRNA generated (MacRae, Zhou et al. 2007). This feature explains that all four Arabidopsis DCLs generate sRNAs of determined and DCL-specific size: 21-nt for AtDCL1 and AtDCL4, 22-nt for AtDCL2, and 24-nt for AtDCL3 (Jacobsen, Running et al. 1999; Golden, Schauer et al. 2002; Schauer, Jacobsen et al. 2002; Xie, Johansen et al. 2004) (Fig.1.9).

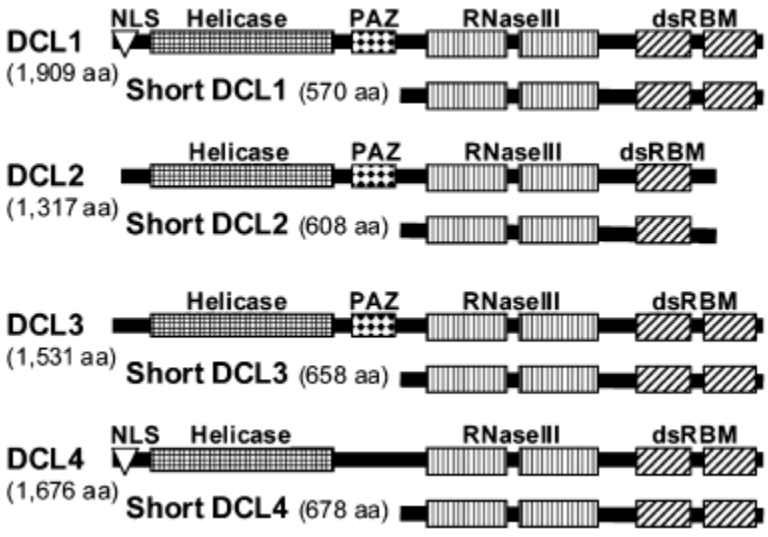


Fig.1.9 - DCLs domains in *Arabidopsis thaliana*
From (Hiraguri, Itoh et al. 2005) modified

In plants the cleavage of the primary transcript to the mature miRNA duplex is thought to be carried out in the nucleus DCL1 (DICER-like protein) (Chen 2005; Jones-Rhoades, Bartel et al. 2006). DCL1 is required for miRNA accumulation, yet processing intermediates do not appear to overacumulate in DCL1 mutants, suggesting that DCL1 has the Drosha like activity responsible for the first set of cuts (Kurihara and Watanabe 2004).

In plants the RNA-binding protein Hyponastic Leaves1 (HYL1) is necessary for pri-miRNA processing (Kurihara, Takashi et al. 2006; Wu, Yu et al. 2007) (Fig.1.10), but may not be the only additional protein involved in processing. In fact, although both HYL1 and DCL1 are required for pri-miR171a processing, they have been identified in protein complexes of significantly different sizes. The HYL1 complex is more or less of 300 kDa (Han, Goud et al. 2004), whereas DCL1 has been identified in a complex of >660 kDa (Qi, Denli et al. 2005).

DCL1 and HYL1 colocalize and are often concentrated in nuclear bodies similar to Cajal bodies (Shaw and Brown 2004; Collier, Pendle et al. 2006; Fang and Spector 2007). The most straightforward interpretation of these observations is that the HYL1 and DCL1 complexes, although not tightly associated, function together in pri-miRNA processing in a distinct nuclear organelle. However, there is not direct evidence that the HYL1-bound miRNA precursors are in the HYL1/DCL1 bodies, maybe because that these nuclear bodies are assembly and storage sites for miRNA processing components, which then function in closer proximity to miRNA gene (Song, Han et al. 2007)

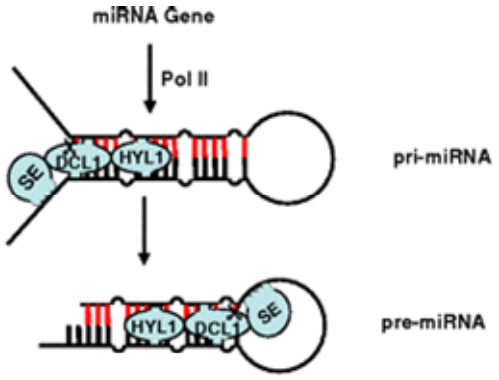


Fig.1.10 - Pri-miRNA processing in plants.
Primary miRNA transcript is processed by the RNaseIII enzyme DCL1 (containing two double-stranded RNA-binding domains) and its associated RNA-binding cofactors HYL1 (containing two double-stranded RNA-binding domains) and SE (a C2H2-type zinc finger) to generate a pre-miRNA. Fom (Zhu 2008) modified.

Several years ago, it was discovered that SERRATE, a C2H2-type zinc finger, is also required for processing pri-miRNAs and for accumulation of mature miRNAs (Grigg, Canales et al. 2005; Yang, Liu et al. 2006) (Fig.1.10).

The *Arabidopsis* gene SERRATE (SE) controls leaf development, meristem activity, inflorescence architecture and developmental phase transition (Prigge and Wagner 2001; Grigg, Canales et al. 2005). SE has also been shown to regulate specific microRNAs (miRNAs), miR165/166, and thus control shoot meristem function and leaf polarity (Grigg, Canales et al. 2005) .

Recently, it was discovered that SE and HYL1 probably act with DCL1 in processing pri-miRNAs before HEN1 in miRNA biogenesis (Yang, Liu et al. 2006) (Fig1.10).

Previous studies have shown that plant miRNAs are processed by DCL1 in a manner similar to animal precursors: a first cut at the base separates the hairpin from the rest of the transcript (Bernstein, Caudy et al. 2001; Kurihara and Watanabe 2004; Vermeulen, Behlen et al. 2005), while a second cleavage releases the mature miRNA (Bologna, Mateos et al. 2009; Schwab and Voinnet 2009)

Recently some miRNAs (ath-miR159 and ath-miR319) were shown to be first processed near the loop, and several 'phased' cuts occur down the stem (Fig.1.11). Mutations and deletions in the upper part of the stem region were shown to abolish processing of these precursors in *Arabidopsis* (Bologna, Mateos et al. 2009; Schwab and Voinnet 2009), while such changes had no effect on the processing of other miRNAs (Mateos, Bologna et al. 2010).

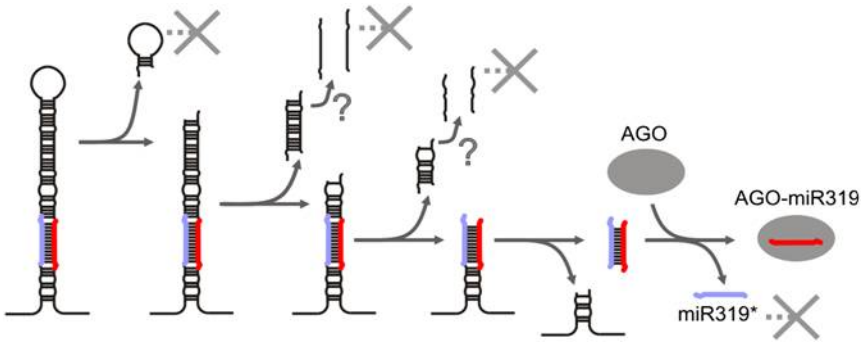


Fig.1.11 - Model of the maturation of *Arabidopsis* miR319 and 159.
From (Bologna, Mateos et al. 2009)

1.4.3 Pre-miRNA processing

Processing is completed by a RNAse III type enzyme the cleaves the pre-miRNA hairpin into the so called miRNA/miRNA* duplex.

In animals the protein responsible for the second cleavage of the precursor miRNA belongs to DICER family and performs the cleavage in the cytoplasm. Instead in plants the cleavege is mediated by the same DCL-1 that processed pri-miRNA into pre-miRNA. Supporting the idea that in plants DCL1 has both Drosha and DICER functions in plant miRNA maturation is the observation that in plants the two sets of cuts that liberate the miRNA/miRNA∗ duplex both occur in the nucleus, which is the predominant location of DCL1.

1.4.3.1 *Pre-miRNA processing in animals*

As mentioned previously, DICER proteins contain: an aminoterminal helicase domain, dual RNAse III motif (RIIIDa that is responsible for the formation of 3'-OH ends of the product and RIIDb that generates the 5' ends), a dsRNA binding domain and a PAZ domain (a 110-amino-acid domain present in proteins like Piwi, Argo and Zwille/Pinheads) (Fig.1.8)

To process pre-miRNA DICER requires partner proteins, such as Loqs in *Drosophila* (Forstemann, Tomari et al. 2005; Saito, Ishizuka et al. 2005), the TAR RNA-binding protein (TRBP) in humans (Chendrimada, Gregory et al. 2005).

This RNAse III type enzyme is among the few nucleases that show specificity for dsRNAs and cleave them with 3' overhangs of 2 to 3 nucleotides and 5'-phosphatate and 3'-hydroxyl terminal. DICER does not recognize just a dsRNA end, but requires its specific structure with a 2-nt 3'-overhang and a 5'-phosphate group. This structure is recognized by the PAZ (Piwi/Argonaute/Zwille) domain which occurs in most DICER proteins. The arrangement of the PAZ domain and RIIID in DICER determines the dsRNA cleavage site (Fig.1.8). The distance between the active center of RIIID and the PAZ pocket accommodating the dsRNA end exactly matches the size of a 25-bp region in an RNA duplex. Bound dsRNA is stretched between the PAZ domain and RIIID of DICER along its flat surface enriched in basic amino acid residues, which interact with the sugar-phosphate backbone of dsRNA. Thus, the size of the resulting miRNA is determined by the distance between the PAZ domain and RIIID (Lee, Ahn et al. 2003; Gregory, Yan et al. 2004; Han, Lee et al. 2004).

The RNase III type enzyme acts as a dimer and thus digests dsRNA with the help of two compound catalytic center domains, with one of them deviating from the consensus catalytic sequence. The crystal structure of the RNAse III catalytic domain was solved and this led to the model for generation of 23 to 28-mer diced miRNA products (Blaszczyk, Tropea et al. 2001). In this model the dimeric DICER folds on the dsRNA substrate to produce four compound catalytic sites so that the two terminal sites bearing partial homology lose functional significance. Thus the DICER product appears to be near the length limit for digestion products of RNAse III enzymes and are double the size of the normal 12- to 15 mer fragments of RNAaseIII enzymes (Yang, Buchholz et al. 2002).

1.4.3.2 *Pre-miRNA processing in plants*

In plants, HYL1 and SE participate in both steps of miRNA biogenesis: from pri-miRNA to premiRNA (see §1.4.2.2) and from pre-miRNA to miRNA, although this is difficult to test vigorously, because the pre-miRNA intermediate does not appear to accumulate and is quickly processed to release mature miRNA. Unlike animal pri-miRNAs, which have an ≈70-nt stem–loop structure where the miRNA is always located ≈11 nt from the base of the stem–loop (Han, Lee et al. 2006), the stem–loop structures of plant pri-miRNAs vary greatly in length (from ≈100 to >1,000 nt) (Sunkar, Girke et al. 2005; Jones-Rhoades, Bartel et al. 2006).

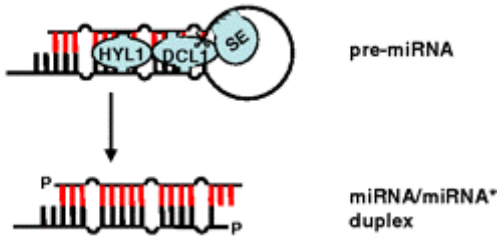


Fig.1.12 - Pre-miRNA processing in plants.
Primary miRNA transcript is processed by the RNaseIII enzyme DCL1 (containing two double-stranded RNA-binding domains) and its associated RNA-binding cofactors HYL1 (containing two double-stranded RNA-binding domains) and SE (a C2H2-type zinc finger) to generate a miRNA/miRNA* duplex.
From (Zhu 2008) modified

In 2008, gel mobility-shift assays performed by Dong *et al.* (Dong, Han et al. 2008) suggest that each of the proteins of the DCL1-HYL1-SE trimeric complex is capable of binding both

pri-miRNAs and pre-miRNAs. In fact it was shown that recombinant HYL1 and SE proteins accelerate the rate of DCL1-mediated cleavage of pre- and pri-miR167b substrates and promote accurate processing (Dong, Han et al. 2008). However, where each protein binds and which sequence and/or structural features they recognize are not known. Crystal structures of each of the proteins bound to a pri-miRNA, and eventually the structures of the trimeric complex with bound pri-miRNA and pre-miRNA will be needed to understand the precise mechanism of accurate miRNA production (Zhu 2008).

In plants, after the cleavage of DCL-1 in the nucleous, the 3' terminal nucleotides of endogenous miRNAs are methylated on their 2' hydroxyl groups by *HUA ENHANCER1* (*HEN1*) (Yu, Yang et al. 2005). Mutations in *HEN1* result in 3' end uridylation of miRNAs which apparently leads to reduced miRNA accumulation and function (Yu, Yang et al. 2005; Yu, Bi et al. 2010). HEN1 contains a methyltransferase domain, and can methylate miRNA/miRNA* duplexes in vitro. The 3' terminal nucleotides of endogenous miRNAs are methylated on their 2' hydroxyl groups in wild-type plants, but not in *hen1* mutants or in animals (Yu, Yang et al. 2005). End-methylation of miRNAs does not enhance silencing activity in vitro and instead appears to protect the 3' ends of silencing RNAs from uridylation and associated destabilization. After DCL1-mediated cleavage and HEN1 mediated methylation, most miRNA molecules exit the nucleus and enter the cytoplasm. This export into the cytoplasm is facilitated by *HASTY* (*HST*), a member of the importin β family of nucleocytoplasmic transporters (Bohnsack, Czaplinski et al. 2004).

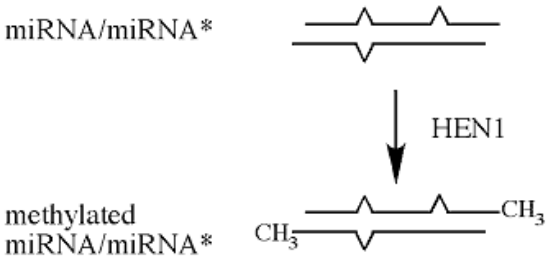


Fig.1.13 - A schematic diagram of miRNA biogenesis in *Arabidopsis*.
The precursor-miRNA (pre-miRNA) is processed by DCL1 to a duplex of the miRNA and its antisense strand miRNA*. HEN1 methylates the 3' terminal nucleotide in each strand of the duplex. The methylation occurs on the ribose of the terminal nucleotide.
Modified from (Yu, Yang et al. 2005)

1.4.4 RISC assembly

Both in animal and in plants, in the cytoplasm, the miRNA/miRNA* duplex is incorporated into a ribonucleoprotein complex, known as the RNA-induced silencing complex (RISC) (Elbashir, Harborth et al. 2001; Hammond, Boettcher et al. 2001)

The RISC has been purified from fly and human cells and in both cases contains a member of the Argonaute protein family, which is thought to be a core component of the complex (Hammond, Boettcher et al. 2001). Argonaute proteins are crucial for RNAi and analogous processes in worms, fungi, and plants, respectively. Argonaute and its homologs are approximately 100 kDa proteins that are sometimes called PPD proteins because they all share the PAZ and PIWI domains (Cerutti, Mian et al. 2000). The PAZ domain (first recognized in Piwi, Argonaute, and Zwille/Pinhead proteins) has a stable fold when isolated from the rest of the protein, which has a barrel core that together with a side appendage appears to bind weakly to single-stranded RNAs at least 5 nt in length and also to double stranded RNA (Lingel, Simon et al. 2003; Song, Liu et al. 2003; Yan, Yan et al. 2003). This dual binding ability suggests that the Argonaute protein could be directly associated with the miRNA before and after it recognizes the mRNA target. When the miRNA strand of the miRNA/miRNA* duplex is loaded into the RISC, the miRNA* appears to be peeled away and degraded.

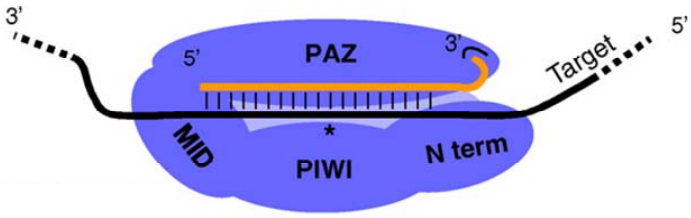


Fig.1.14 - Argonaute domains: MID domain, PIWI, PAZ.
From (Vazquez, Legrand et al. 2010) modified

Loading sRNAs into AGO effectors involves selecting one strand as the guide for target identification and removing the passenger strand (or star strand noted miRNA* or siRNA*).

In animals, the strand with the least stable 5'-end is selected as the guide (Hutvagner 2005). Although a contribution of the 5' stability is still debated for plant sRNAs, some AGOs have been shown to select the guide strand depending, at least in part, on the identity of its first 5'-nucleotide through interaction with a nucleotide-specific binding pocket located in the MID (MIDDLE) domain (Rajagopalan, Vaucheret et al. 2006; Mi, Cai et al. 2008;

Montgomery, Howell et al. 2008; Takeda, Iwasaki et al. 2008; Vaucheret 2008; Eamens, Smith et al. 2009). The 2-nt 3'-overhang of the selected strand might then be stabilized by anchoring to the PAZ domain, as suggested for Drosophila AGO2 (Song, Liu et al. 2003). It is unknown how plant AGOs remove the passenger strands to allow guide-strand pairing with the target, i.e. whether this occurs as for mice AGO2 by cleavage through the 'slicer' activity (Song, Liu et al. 2003; Chendrimada, Gregory et al. 2005; Vaucheret 2008) located in the PIWI domain or like for Drosophila AGO1 by passive unwinding (Rand, Petersen et al. 2005; Kawamata, Seitz et al. 2009).

1.5 MiRNA targets

The importance of complementarity to the 5′ terminal of metazoan miRNAs has been suspected since the observation that the *lin-14* UTR has "core elements" of complementarity to the 5′ region of the *lin-4* miRNA (Wightman, Ha et al. 1993).

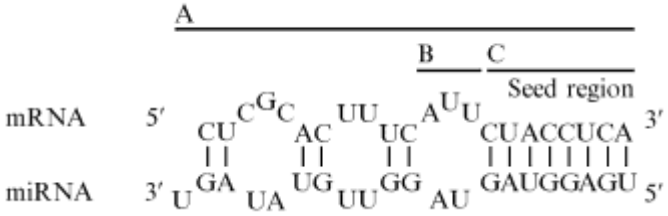


Fig.1.15 - Typical pattern of base pairing between miRNAs and target.
A= miRNA/mRNA duplex region; B=central region of base pairing;C= 5' end of the miRNA (seed). According to thermodynamic analysis some degree of complex formation occurs along the entire miRNA—mRNA region.Usually the interaction is weak in B and strong in C.

More recent observations support this idea:

(1) Residues 2–8 of several invertebrate miRNAs are perfectly complementary to 3′ UTR elements previously shown to mediate posttranscriptional repression. (Lai 2002)This region of perfect complementarity is called the "seed" region in animal miRNAs. In plants there is not the seed, because the complementarity with the target is perfect or near perfect for the length of the miRNA.

(2) Residues that pair to residues 2–8 of the miRNA of invertebrates are usually perfectly conserved in orthologous transcripts of other species (Stark, Brennecke et al. 2003).

(3) Residues 2–8 of the miRNA are the most conserved among homologous metazoan miRNAs (Lewis, Shih et al. 2003; Lim, Lau et al. 2003).
(4) The perfect pairing of mRNA to the residues 2–8 of the miRNA (seed region) (Fig.1.15). is much more productive than pairings to any casual heptamer of the miRNA (Lewis, Shih et al. 2003). Pairing to this 5′ core region also appears to disproportionally regulate the specificity of siRNA-mediated mRNA cleavage (Jackson, Bartz et al. 2003; Pusch, Boden et al. 2003) and the same happens in plant miRNAs that generally mediate transcript cleavage.

1.5.1 Approaches for prediction and validation of miRNA targets in animals

Significant problems beset bioinformatics approaches to target identification in animals: the complementarity between miRNAs and target mRNAs is usually imperfect: only the short region the 'core' (seed) at the 5′ side of the miRNA is perfectly base-paired to the transcript (Fig.1.15). Consequently, a search for sequence complementarity will produce many results and many false-positive predictions. Moreover, most miRNAs have several targets, some of which will be targeted more strongly than others. Ideally, one should therefore predict not just targets, but also the expected degree of translational suppression. To overcome this problem, was developed software, such as PicTar (Grun, Wang et al. 2005) and TargetScan (Lewis, Burge et al. 2005), that take into account the evolutionary conservation of the target site. Although conservation is a powerful way to improve the detection signal, it is clearly not useful for the lineage-specific miRNAs (Bentwich, Avniel et al. 2005).
A limitation of these approaches is that they assign the same score to all targets with the same sequence and cannot explain the variability that arises from differences in accessibility imposed by the sequence surrounding the target. Target accessibility is a critical factor in microRNA function and several studies have shown that miRNA target sequences tend not to be involved in energetically stable cis-secondary structure elements (Kertesz, Iovino et al. 2007; Long, Chan et al. 2008). In fact recent study about the conformational modification of mRNA for allowing the interaction with the miRNAs pairing suggested to Kertesz (Kertesz, Iovino et al. 2007)to account for the effect of accessibility on the strength of microRNA repression. In this study, an energy-based score for microRNA-target interactions, DDG, equal to the difference between the free energy gained by the binding of the microRNA to the Target (DGduplex) and the free energy lost by unpairing the target-site nucleotides was calculated (DGopen) (Kertesz, Iovino et al. 2007).

Other target prediction algorithms have been proposed for animals (eg: (Wang 2006; Sturm, Hackenberg et al. 2010)). Due to the nature of animal miRNA/target interactions, it is generally considered that these methods are likely to be able to identify most real targets, but might tend to also yield a large proportion of false positive target predictions.

Experimental validation of target mRNAs in animals is hard because interaction of an mRNA with a target (and sequestration from the transcriptional apparatus) does not necessarily affect steady state mRNA levels, protein expression must therefore be followed, preferably in conjunction with approaches to show inverse correlation with miRNA expression.

1.5.2 Approaches for the prediction and validation of miRNA targets in plants

The fact that most plant microRNAs are thought to form almost perfect hybrids with their target sequences (which are typically situated in coding regions) significantly aids target prediction. Most approaches are therefore based on the search for sequences showing extensive complementarity to a known miRNA. However, experimental studies have suggested that certain types of mismatch at certain positions in a miRNA/target hybrid are unlikely to be compatible with target cleavage. Accordingly, a widely used scoring system for mismatches was developed by Allen et al (Allen, Xie et al. 2005). Additionally, recent studies strongly indicate that a significant proportion of plant miRNAs are likely to form target interactions that resemble those in animals and might lead to translational sequestration rather than message degadation (Lanet, Delannoy et al. 2009)

At present, the most widespread method to confirm functional miRNA-mRNA targets in plants, where such interactions typically result in target mRNA cleavage is the 5′ RACE procedure (Random Amplification of cDNA Ends).

5′ RACE has been used by many researchers to identify miRNA targets in plants (Palatnik, Allen et al. 2003; Mallory, Reinhart et al. 2004; Mallory, Bartel et al. 2005; Sunkar, Girke et al. 2005). Cleaved mRNA products in plants have two properties:

- the 5′ phosphate of a cleaved mRNA product can be ligated to an RNA adaptor with T4 RNA ligase;
- the precise target cleavage position is the mRNA target nucleotide pairing with the tenth nucleotide of miRNA (Sunkar, Girke et al. 2005);

- mRNA cleavage products of miRNA guided activity can be amplified with ligation of an oligo-nucleotide adaptor to the 5′ end, followed by reverse transcription and PCR amplification with a gene specific primer.

A modified 5′ RACE procedure can be applied as follows. Total RNA is isolated and polyA mRNA is prepared and directly ligated to an RNA oligo adaptor. Oligo dT is used to synthesize the first strand of cDNA with reverse transcriptase. This first cDNA strand is amplified non-specifically. Then the 5′ RACE PCR and 5′ nested PCR are performed using primers specific to the adapter sequence and to the gene predicted to be targeted by a given miRNA. RACE products are gel purified, cloned, and sequenced.

More recently, several groups have proposed a high throughput approach known as Degradome (Addo-Quaye, Eshoo et al. 2008) or Parallel Analysis of RNA Ends (PARE) (German, Pillay et al. 2008), whereby the 5' 20 bases of the downstream fragment of RNAs degraded by RISC can be sequenced in a massively parallel manner, effectively giving a snapshot of all degraded mRNAs in a cell or tissue. Bioinformatics analysis allows identification of which messages are represented and at which positions they are cleaved. Correspondence between over-represented cleavage sites and predicted miRNA targets can allow the large scale validation of miRNA targets. This method can be simply visualized as a manifestation of the 5'-RACE approach where all - or at least most - target genes are studied simultaneously.

1.6 MiRNA prediction, validation and quantification

1.6.1 Conservation and evolutionary aspects – comparative prediction

Many, but by no means all mature microRNAs are conserved exactly or closely between groups of phylogenetically related organisms (and can be produced from more than one locus in the same organism). Where detailed experimental characterizations of miRNA target interactions have been performed, it is clear that interactions between such miRNAs and their targets are also conserved. Such groups of related miRNAs are considered to be part of the same miRNA family. Indeed, while the degree of sequence conservation is highest between the mature miRNAs within families, in many cases, sequence similarity between precursor loci confirms homology of such loci. Around 28 families of miRNAs are known to be deeply conserved within plants (present in at least 2 monocot organisms and at least 2 dicots). Many other families are conserved within monocots or within dicots. A

similar situation exists within animals although conclusive evidence of conservation of miRNA families between plants and animals is elusive.

In any case, the principle of conservation of miRNA families has been used as the basis for a series of so called "comparative genomic miRNA prediction tools". The main idea behind this approach is finding potential homologs of known mature miRNAs or pre-miRNA sequences either within a single genome or between genomes of related organisms (Lagos-Quintana, Rauhut et al. 2001; Lee and Ambros 2001). Regions of showing similarity to known miRNAs are either scanned for pre-miRNA-like structures, or alignments are examined for traces of selective constraints favouring conservation of miRNA-like traits.
The softwares miRscan (Lim, Lau et al. 2003), miRseeker (Lai, Tomancak et al. 2003) and miralign2 (Wang, Zhang et al. 2005)contributed to the prediction of many novel microRNAs in nematodes (Lim, Lau et al. 2003), insects (Lai, Tomancak et al. 2003; Wang, Zhang et al. 2005)and vertebrates(Lim, Glasner et al. 2003). In 2003, Grad developed a method for predicting miRNAs in the nematode genome of *C. elegans* using both sequence and structure homology and comparing it with known miRNAs (Grad, Aach et al. 2003). A similar approach was employed by Dezulian for plant miRNAs prediction (Dezulian, Remmert et al. 2006) and Jones-Rhoades and Bartel(Jones-Rhoades and Bartel 2004). In 2005, Berezikov used phylogenetic shadowing to find regions that are under stabilizing selection and exhibit the characteristic variations in sequence conservation between loop, stem and mature miRNA(Berezikov, Guryev et al. 2005). In this case, secondary structure is used in a later filtering step.
Genomic context also can give additional information: Mirscan-II, for example, takes conservation of surrounding genes into account (Ohler, Yekta et al. 2004) In 2005 Altuvia utilize the propensity of miRNAs to appear in genomic clusters (often in the form of polycistronic transcripts) as an additional selection criterion(Altuvia, Landgraf et al. 2005).

Work presented in chapter 2 of this thesis concerns comparative prediction and validation of miRNAs in the grapevine, *Vitis vinifera.*

1.6.2 *Ab-initio* prediction of miRNAs

While many of the aforementioned approaches rely on knowledge of a candidate mature miRNA sequence conserved between genomes, others – principally those that focus

on structural aspects of entire hairpins rather than the candidate miRNA/miRNA* region - can be used to seek novel pre-miRNA sequences, often in regions showing conservation between related genomes (eg, (Berezikov, Guryev et al. 2005)). The bioinformatics search for novel miRNAs which are not part of previously known families is known as *ab-initio* miRNA prediction. Such methods, in the absence of a-priori information on the sequence of the mature miRNA, must consider general characteristics of hairpins to discriminate between real pre-miRNAs and non-miRNA hairpins. For genome-wide predictions, all thermodynamically stable potential hairpins are predicted from RNA structure modelling approaches and submitted to an algorithm to evaluate the probability that structures could correspond to real pre-miRNAs. To avoid large numbers of false positive predictions, such *ab-initio* methods must have very high specificity of prediction as large genomes can contain many millions of potential hairpin structures.

The miR-abela3 approach first searches for hairpins that are very stable against changes in the folding windows and then uses a support vector machine (SVM) to identify microRNAs among these candidates (Sewer, Paul et al. 2005). A related technique is described by Xue et al. (2005) that use as input for the SVM features related to the frequency of triplet of nucleotides(Xue, Li et al. 2005). The program PalGrade scores hairpins in a similar way (Bentwich, Avniel et al. 2005).

A quite different approach starts with the analysis of overrepresented patterns in phylogenetic footprints located in the 3'UTRs of mRNAs. These motifs constitute putative microRNA target sites and are used to guide the search for corresponding pre-miRNA candidates (Xie, Allen et al. 2005). Other *ab-initio* predictors have utilized Context Specific Hidden Markov Models (Agarwal, Vaz et al. 2010), Genetic Programming, (Brameier and Wiuf 2007) the identification of context robust hairpins physically close to known animal pre-miRNAs (Sewer, Paul et al. 2005).

Work presented in chapter 3 of this thesis concerns the development of *ab-initio* pre-miRNA prediction methods

1.6.3 Validation/quantification of miRNA predictions

In principle, it might be possible to experimentally validate the in vivo presence of any of the three main stages of miRNA biogenesis: pri-miRNA transcripts, precursor

hairpins or mature miRNA sequences. In practice, the physiological half life of pri-miRNAs and pre-miRNAs is generally rather short. Several studies have identified some pri-miRNA transcripts either in EST/cDNA collections (Hubbard, Grafham et al. 2005; Mica, Piccolo et al. 2010), or, where primary microRNA transcripts are poly adenylated, through conventional RACE strategies(Szarzynska, Sobkowiak et al. 2009). This strategy is especially effective if mutants in genes essential for miRNA processing are available as they can lead to the accumulation of primary transcripts.

However, the vast majority of experimental validations of miRNAs are performed by testing for the presence of the mature miRNA through RT-PCR (Bandres, Cubedo et al. 2006; Liu, Fan et al. 2009), northern blotting (He, Nie et al. 2008; Zhao, Yang et al. 2009), or oligonucleotide array experiments (He, Nie et al. 2008; Zhao, Yang et al. 2009; Mica, Piccolo et al. 2010). In the same spirit, quantification of differential expression of miRNAs between tissues, developmental stages or experimental conditions is routinely performed by real-time PCR, northern blotting, oligonucleotide array or, more recently, by quantitative analysis of smallRNA deep-sequencing data (Liang, Zhang et al. 2010; Mica, Piccolo et al. 2010; Zhao, Xia et al. 2010).

1.6.4 Deep sequencing and bioinformatics

A more direct approach to the discovery of miRNAs is to isolate and sequence smallRNAs themselves (mature miRNAs and other smallRNAs), and to then map them to the genome of origin, checking that they fall in stem regions or hairpins. Usually, bioinformatics tools, similar to those used in comparative or *ab-initio* miRNA discovery are used to evaluate the possibility that any given locus that generates smallRNAs is a microRNA precursor, rather than the smallRNA being produced by other biogenesis methods.

Initially, such experiments were performed using conventional Sanger di-deoxy sequencing, often of concatenated series of smallRNAs. More recently the advent of so-called Next Generation Sequencing strategies such as the Roche 454 (Margulies, Egholm et al. 2005; Berezikov, Thuemmler et al. 2006), ABI SOLiD (Ribeiro-dos-Santos, Khayat et al. 2010) and Illumina Genome Analyser platforms (Berezikov, Thuemmler et al. 2006; Denoeud, Aury et al. 2008), has revolutionized discovery of miRNAs by sequencing. These technologies are capable of reading tens of millions of complete smallRNA sequences in a

single reaction, and thus, relatively deep coverage of expressed smallRNAs can be easily obtained. These methods are essentially quantitative in their coverage, meaning that relative expression levels of different miRNAs can be estimated from the frequency with which different sequences are observed. Reactions are often performed for different tissues, developmental stages or experimental conditions.

In fact, the main difficulty in the analysis of such data is that they require dedicated bioinformatics tools to efficiently map the loci of origin of reads on the genome sequence, and to perform extremely high thoughput analysis of secondary structures predicted. Such methods have recently become the main strategy for miRNA discovery and quantification in both animal and plant systems (Batuwita and Palade 2009; Legeai, Rizk et al. 2010; Liang, Zhang et al. 2010; Mica, Piccolo et al. 2010).

Work presented in chapter 4 of this thesis concerns the use of deep sequencing of smallRNA to identify novel miRNAs in the grapevine, *Vitis vinifera*

1.7 Thesis background and structure

This thesis contains work performed as part of a large collaborative effort aimed principally at the characterization of the genome of the grapevine, *Vitis vinifera*, which was sequenced, assembled and annotated by the French Italian Public Consortium for Grapevine Genome Characterization (Jaillon, Aury et al. 2007). In particular, the work presented here focuses on the implementation of different bioinformatics strategies for the identification of both conserved and novel miRNAs in this species. The first results chapter describes comparative prediction of conserved miRNAs in *Vitis* and the use of high throughput methods to validate their expression. The following chapter concerns the implementation of an *ab-initio* pre-miRNA prediction tool, while the final chapter outlines novel strategies for the interpretation of smallRNA deep sequence data and their application to the discovery of miRNAs in the grapevine genome.

Chapter 2

An edited version of this chapter was published as:

Mica E, Piccolo V, Delledonne M, Ferrarini A, Pezzotti M, Casati C, Del Fabbro C, Valle G, Policriti A, Morgante M, Pesole G, Pè ME, Horner DS (2010). High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in *Vitis vinifera*. *BMC Genomics*, **11**:109

2. Comparative prediction of grapevine miRNAs

2.1 Introduction

In august 2007, the French-Italian sequencing project released the first high quality draft of the *Vitis vinifera* genome sequence (Jaillon, Aury et al. 2007).

Our contribution to the project was the prediction of small RNAs in grapevine genome (tRNAs, snoRNAs, miRNAs, srpRNAs, rRNAs). Non miRNA predictions were performed with standard tools and the results and discussion of these data lies outside the scope of this thesis.

As discussed previously, homologs of members of conserved miRNA families are often annotated in newly sequenced genomes through comparative methods. Essentially, such methods search for short regions showing sequence similarity to known mature miRNAs and then examine the potential of the flanking regions to assume secondary structures compatible with their being miRNA precursors. Such approaches have been demonstrated to be remarkably effective even if they are unable to identify novel or *lineage specific* miRNAs.

For the initial predictions we chose a comparative approach able to detect homologs of known miRNAs in a newly sequenced genome. The computational tool we used was MicroHARVESTER software (Dezulian, Remmert et al. 2006) that searches for miRNA homologs in one or more query sequences.

2.2 Comparative prediction of miRNAs in grapevine using microHARVESTER

Given a known miRNA (miRNA precursor sequence plus mature miRNA sequence) as input, microHARVESTER uses the precursor as a query for a sequence similarity search against a set of sequences the genome under study to generate a set of candidate homologs. Since mature miRNAs are often highly conserved (Axtell and Bartel 2005) using BLAST (Altschul, Gish et al. 1990) with a very high *E*-value cutoff and minimal word size of 7,

generates hits for almost all miRNA homologs at the price of many false positives. MicroHARVESTER then applies a series of filters to remove poor candidates. First, candidates whose aligned segments do not span most of the mature miRNA part of the query are discarded. In a second filter step, a modified Smith–Waterman pairwise alignment algorithm (Smith and Waterman 1981) is used to precisely determine the mature sequence in the candidate precursor from the optimal alignment of the query mature sequence against the corresponding segment of the BLAST hit. Candidates where the length of the mature sequences differs by >2 nt from the expected length are discarded. In a third filter step, the minimal free energy structure of the candidate sequence is predicted using RNAfold (Hofacker, Fontana et al. 1994; Hofacker 2003; Hofacker, Bernhart et al. 2004) and the putative miRNA* sequence is determined. Candidates are discarded if more than six nucleotides of its miRNA* are not predicted to form bonds with its mature miRNA. We created a database of all plant miRNAs present in the present in release 9.1 of miRBase (Griffiths-Jones, Saini et al. 2008) and then we searched for homologs in the entire grapevine genome. 140 high confidence predictions were generated (Jaillon, Aury et al. 2007).

In late 2009, a second draft of the grapevine genome, based on 12x sequencing coverage was released. Repetition of the microHARVESTER analysis revealed that 2 loci originally included in the 8x genome were no longer present in the new assembly and probably represented fragments that had been included twice in the intitial draft. However, 10 new loci corresponding both to additional loci from known families and members of families not previously identified in *Vitis* were identified.

In our analysis we confirm existing patterns of miRNA family conservation with respect phylogenetic distribution of miRNAs annotated in miRBase.

Of the 30 families for which we identified putative precursor sequences in *Vitis vinifera*:

- 26 are known to be deeply conserved (present in monocots and dicots) ;
- 3 are families thought to be specific to dicots (miR403, 477, 479)

In fact, we detected members of 26 of the 27 families that are known to be deeply conserved (miRBase contains 36 families proposed to be deeply conserved, although 9 of these are poor predictions and annotated as dubious in the database). We did not find putative homologs of any of these 9 (miR413, 414, 415, 416, 417, 418, 419, 420, 426). The sole confirmed deeply conserved family for which microHARVESTER did not find a locus in *Vitis* is miR2118, which was not known at the time the analyses was performed and which we subsequently identified through an alternative approach (see Chapter 4).

Given that representatives of all expected families and some additional families, previously identified in at least one dicot) were found lead us to conclude that the microHARVESTER analysis generated adequate results in the identification of conserved miRNA families.

We mapped all miRNAs precursor with respect to the reference annotation of protein coding genes in the grape genome:

- 134 pre-miRNAs were intergenic in location. In particular 17 precursors overlapped with annotated genes but on the non-coding strand (opposite strand respect to the gene).
- 4 precursor predictions fell within or overlapped annotated coding or UTR exons although homology searches and transcriptomics data generated subsequently to the initial annotations call into question the validity of all but two of these exon annotations. We noted that miRNA 156 h is probably an incorrect prediction derived from a coincidentally plausible hairpin structure formed by the opposite strand to the presumed target (a *Squamosa-promoter Binding Protein* (*SBP*) box gene). A similar situation is observed for miR171g which falls on the opposite strand to a GRAS domain transcription factor gene. These predictions were removed from our candidate set.
- 9 precursor were apparently intronic in location. We controlled them manually and noted that all of the introns putatively containing pre-miRNAs were likely to be erroneous predictions, being atypically long (over 13 kb) and interrupting putative retroelement derived genes or obvious fusion gene predictions (not shown).

These patterns conform to those expected for plant miRNAs in that very few such sequences have been shown to be intronic in location.

The coordinates, mature mirRNA sequences and other information regarding subsequent analyses of the 146 conserved miRNA loci identified on the 12x assembly and retained in our final prediction set are presented in Tab.2.1.

2.3 Validation of expression pattern of mature miRNAs using oligonucleotide arrays

2.3.1 Evaluation of *Vitis vinifera* conserved miRNAs using oligonucleotide array technology

Expression patterns and levels of miRNAs for which the sequence is known can be evaluated using experimental approaches such as northern blot or real-time PCR. However,

these methods are laborious and time consuming when applied to large numbers of candidates.

Currently, one of the preferred high throughput technologies to check tissue expression specificity is the oligonucleotide array, where an arrayed series of thousands of microscopic spots of DNA oligonucleotides, complementary to the expected miRNAs, are anchored to a slide and hybridized to labelled smallRNA isolated from tissue samples. Intensity of labelled RNA hybridized to individual probes are measured with high precision optical instruments and reflect expression levels of individual miRNAs (Davison, Johnson et al. 2006; Liu, Calin et al. 2008; Liu, Spizzo et al. 2008; Yin, Zhao et al. 2008). As with conventional protein-coding genes, it is of interest to profile the expression patterns of microRNAs in many tissues during different conditions (e.j. development, stress, disease).

In 2008 in collaboration with Erica Mica and Prof. Enrico Pè (of the University of Milan and more recently the Scuola Superiore Sant'Anna, Pisa) and Prof. Mario Pezzotti and Prof. Massimo Delledonne (University of Verona) we studied the expression profile of the predicted microRNAs using a CombiMatrix 12 K CustomArray platform (Mica, Piccolo et al. 2010). Advantages of this platform with respect to some others include the possibility to quickly and cheaply generate custom oligonucleotide arrays and to analyse simultaneously expression patterns in more than one tissue (the array is divided into different compartments that can be hybridized simultaneously to different RNA samples). In this way was possible to check if the conserved microRNAs, that we predicted in grapevine show differential expression patterns between tissues and during the maturation of fruit (Mica, Piccolo et al. 2010) We used as tissues three stages of ripening berries (immature berry, *veraison*, mature berry), roots, leaves and young inflorescences.

In order to demonstrate specificity of hybridization and to show that resulting signals indeed derived from mature miRNAs and not from precursor sequences or non-specific interactions, it is necessary, for each miRNA, tested to include probes complementary to the mature miRNA sequence and also probes to regions of the precursor that are not expected to be present after pre-miRNA processing (sequences derived from the loop region). Small amounts of miRNA* sequences are expected to be present and so probes to these sequences are included. In order to confirm that smallRNAs detected are produced in a strand specific manner, it is also desirable to include reverse complementary probes to the miRNA and miRNA* sequences. Furthermore, to probes shifted by a few bases with respect to the expected mature miRNA and miRNA* sequences can assist in the statistical analysis of hybridization specificity as can probes where several destabilizing substitutions are

introduced into the sequences. Of course positive and negative controls are also included in such strategies.
Our task in this collaboration was the design of the microarray chip

2.3.2 Bioinformatics approach to oligonucleotide array design.

The genomic sequence of a putative pre-miRNA structure was extracted and the global most stable secondary structure was estimated using RNAFold (Mccaskill 1990; Hofacker, Fontana et al. 1994; Hofacker and Stadler 2006; Mica, Piccolo et al. 2010). RNAfold exports a textual representation of the predicted secondary structure where each bracket corresponds to a base-pairing, in particular "(" pairs with ")", while each dot "." corresponds to a base not paired (it could be mismatch or a base in excess). For example loops are represented uniquely by dots, while stems contain a variety of base parings, bubbles, mismatches (Fig.2.1).

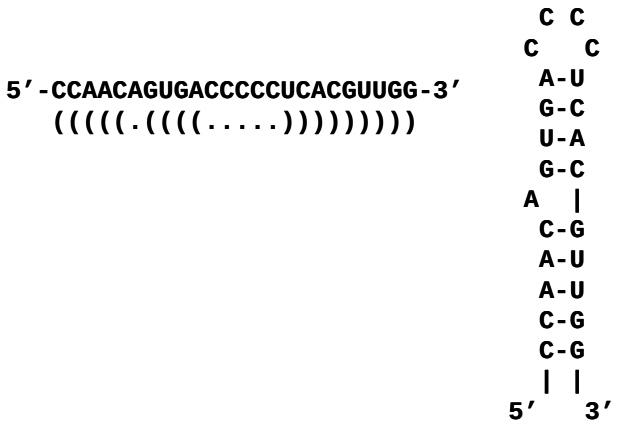


Fig. 2.1 - Textual rapresentation of secondary structure, according to RNAfold.
Each bracket corresponds to a base-pairing, in particular "(" pairs with ")", while each dot "." corresponds to a base not paired (it could be mismatch or a base in excess).

In order to allow rapid evaluation of predicted secondary structures, we developed a script in PERL (http://www.perl.org/) able to scan all base pairing possibilities using the textual representation of secondary structure and to accurately identify the expected miRNA* sequence for any specified mature miRNA.
The script we developed is articulated and complex. It scans the hairpins starting simultaneously from the 5' and from the 3' termini, finding all correct base pairings (defined

as "(" and ")"), and all mismatches (represented as positionally corresponding pairs of ".") or all bubbles (defined as one or more dot "." without corresponding bases in the opposite arm). If we know the mature miRNA sequence and want to detect its correspondent miRNA*, we first should map the sequence of mature into the secondary structure and then shift two nucleotides in order to find the start position of the miRNA* according to the known patterns of dsRNA cutting performed by enzymes of the DICER family (Fig.2.2) (Bernstein, Caudy et al. 2001; Lee, Kim et al. 2004).

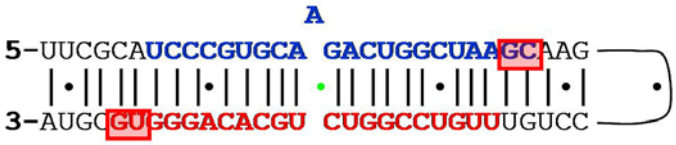


Fig. **2.2 - Example of *overhangs* on the 3' end of the pre-miRNA sequence**
We use the characteristic of 2-nucleotide *3'* overhang in order to detect the miRNA* sequence of each mature miRNA

For each grapevine miRNA precursor, we designed a set of 20-22 nt probes specific for:

- the mature miRNA;
- the correspondent miRNA*;
- mature miRNA and miRNA* complementary sequences;
- probes shifted 5 or 10 bases 3' or 5' with respect to the central base of the corresponding mature miRNAs;
- probes derived from regions of the stem (not predicted to overlap with the mature miRNA sequence) and from the loop.

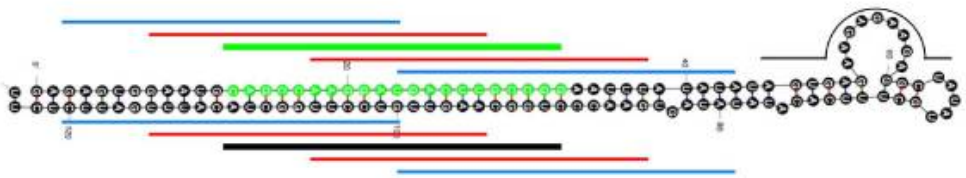


Fig.2.3 - Oligonucleotide design strategy for Combimatrix custom oligonucleotide array.
Probes were designed complementary to the predicted mature miRNA (green line) and miRNA* (thick black line) sequences. Additional probes were designed to the loop region (thin black line) as well as probes shifted 5 nucleotides (red lines) and 10 nucleotides (blue lines) with respect to the miRNA and miRNA* sequences.
From (Mica, Piccolo et al. 2010)

As positive controls we used snRNA U6 and four grapevine tRNA. In addition, we added other fourteen distinct negative and mRNA degradation control probes. Additionally, for each specific probe, a mismatch control with 2 maximally destabilizing substitutions was included. Each probe was present on the final array in three replicates.

Slides were hybridized with 3 micrograms of LMW RNA labeled with Cy5 (Mirus LabelIT miRNA labeling Kit (Mirus Bio Corp.)). Hybridization and washing were performed as indicated by CombiMatrix. Slides were scanned with a Perkin Elmer Scanarray 4000 XL raw data was extracted with Scanarray Express 4.0 and Microarray Imager (CombiMatrix) software. After each hybridization, slides were stripped according to manufacturer's instructions and re-used 5 to 6 times.

Two hybridizations were performed with independently extracted LMW RNAs, for each sample. Background level was defined as the average signal of the negative and degradation controls plus two times their standard deviation. The ratio between intensities of the perfect match probe and its mismatch probe (referred to as PM/MM) was also used to estimate the reliability of each signal. Probes with a median signal higher than background and with PM/MM value higher than 1.2 were called as present. The normalization between arrays was performed using the quantile normalization method (Bolstad, Irizarry et al. 2003) using the BLIST software, provided by Combimatrix.Normalized signals were Log2 transformed and probes with a low PM/MM ratio (<1.2) were discarded. Differentially expressed genes in various tissues were identified with a one-way ANOVA test (p-value < 0.05). Significant results were further investigated with Scheffè test, a *post hoc* test to define which tissues showed significant differences.

RNA extractions were performed by Dot.ssa Erica Mica and array synthesis, hybridizations and signal analysis were performed in the labs of Prof. Mario Pezzotti and Prof. Massimo Delledonne of the University of Verona. The majority of statistical analyses of data generated were performed by Dot.ssa Erica Mica.

2.3.3 Results of oligonucleotide array analyses.

Of the mature miRNA sequences considered, 56 (corresponding to 23 different families), showed significant expression in at least one tissue tested (Tab.2.1) and another 6 showed a borderline signal. Specifically, 41 different miRNAs showed significant signal in roots, 47 in leaves, 49 in young inflorescences, 53 in green berries, 42 in berries at *veraison* (the point where growth ends and maturation begins) and 40 in mature berries.

We didn't find a significant hybridization for mismatch and shifted probes, except for probes shifted 5 nucleotides towards the 5' end of the miRNA precursor. In fact for more than 90% of the probes a signal drop-off greater than 90% was observed between the miRNA probe and shifted probes.

For probes shifted 5 nucleotides towards the 5' end of the miRNA precursor the lack of signal drop-off might be due to the fact that probes were synthesized with their 3' termini towards the slide, and that no "spacer" oligonucleotide was used (according to CombiMatrix protocols). As a consequence, steric effects might reduce the specificity determined by the 3'-most five bases of the probes.

Other than for 26 out of 140 pre-miRNAs, no detectable signals were recorded for the probes designed on the precursor loop regions - likely due to size fractionation of RNA samples and the relatively short half-life of pre-miRNAs. We can conclude that our miRNA expression data are principally derived from mature miRNAs molecules, without appreciable pre-miRNA contamination.

Finally, it should be noted that recent studies have demonstrated appreciable levels of cross-hybridization between closely related miRNAs and probes differing by only one or two bases (Barad, Meiri et al. 2004). It is therefore difficult to exclude the possibility that cross-hybridization within miRNA families causes a distortion of quantitative estimates of expression levels of some individual mature miRNA sequences.

Of the mature miRNA sequences considered, 56 (corresponding to 23 different families), showed significant expression in at least one tissue tested, and another 6 showed a borderline signal. Specifically, 41 different miRNAs showed significant signal in roots, 47 in leaves, 49 in young inflorescences, 53 in green berries, 42 in berries at veraison (the point where growth ends and maturation begins) and 40 in mature berries.

To evaluate the statistical significance of the differential expression of mature miRNAs in the six tissues considered, we set up two distinct comparisons: one among the three developmental stages of the ripening berries and the other one among leaves, roots and inflorescences. ANOVA analyses were performed with a P-value threshold of 0.05 and subsequently a Scheffè test was used to assess which of the three tissues showed significant differences. Thirteen different mature miRNAs showed a statistically significant change in signal between the ripening stages of the berry (Fig. 2A-C), and 27 miRNAs showed significant changes in their expression when comparing three different tissues (leaves, roots and inflorescences)(Fig.2.4.D-H). miR395a and miR171h show a distinctive pattern of expression - being highly expressed at veraison with respect to the other two stages (4.4 and

2.3 fold changes of expression level respectively) (Fig.2.4.A). Seven miRNAs (miR156f, miR169a, miR169f, miR169r, miR169x, miR319b and miR535a) are more expressed in mature berries than in green berries (Fig.2.4.B). Four miRNAs (miR171c, miR172c, miR396c, miR403a) are, on the contrary, more expressed in green berries, their expression decreasing during ripening (Fig.2.4.C).

Clear patterns also emerge from analyses of differential expression between roots, leaves and young inflorescences. Thirteen miRNAs are significantly differentially expressed in roots, showing a similar expression in the other tissues. In particular miR397a, miR398b and miR408 all show at least 100 fold higher expression in root than either leaf or early inflorescences, while miR159a, miR160a, miR399a, miR399b, miR403a and miR535 show more modest, but still significant, changes in the same comparisons (Fig.2.4.D). On the contrary miR164a, miR164b, miR171c and miR172c show a significantly lower level of expression in roots (Fig.2.4F).

Five miRNAs (miR169v, miR169y, miR171f, miR171h and miR319b) yield significantly higher signals in young inflorescences than both leaves and roots (between 2 and 7.2 fold higher levels in this tissue) (Fig.2.4E). Only one miRNA, miR160c, shows a leaf-specific expression profile (2.5 fold lower level in leaves with respect to other tissues) (Fig.2.4G). Finally, six miRNAs (miR169a, miR169e, miR169f, miR169x, miR171e and miR395a) exhibit significant differences in expression levels in all comparisons between leaf, root and inflorescences (Fig.2.4H). Five of these miRNAs (169a, 169e, 169f, 169x and 171e) show the highest expression in young inflorescences and the lowest in roots.

Following the widespread assumption that many miRNA/target interactions are conserved between related species (Bartel 2004; Jones-Rhoades, Bartel et al. 2006) our data regarding differential expression of mature miRNA sequences raise some intriguing possibilities particularly with respect to the potential importance of miRNA in the regulation of fruit maturation.

Li et al. (Li, Oono et al. 2008) recently showed that the transcription factor NFYA5 is targeted by miR169 and that overexpression of miR169 leads to excessive water loss through leaves and hypersensitivity to drought stress in *Arabidopsis. thaliana*. In this light, the preponderance of miR169 family members in the group of miRNAs upregulated in mature berries is striking and might reflect a mechanism to protect maturing fruit from dehydration.

We also note that miR535 family, identified so far only in *O. sativa* and *P. patens (Arazi, Talmor-Neiman et al. 2005)* is upregulated during berry maturation. This is a first

indication of a possible function of miR535 for which no information was previously available.

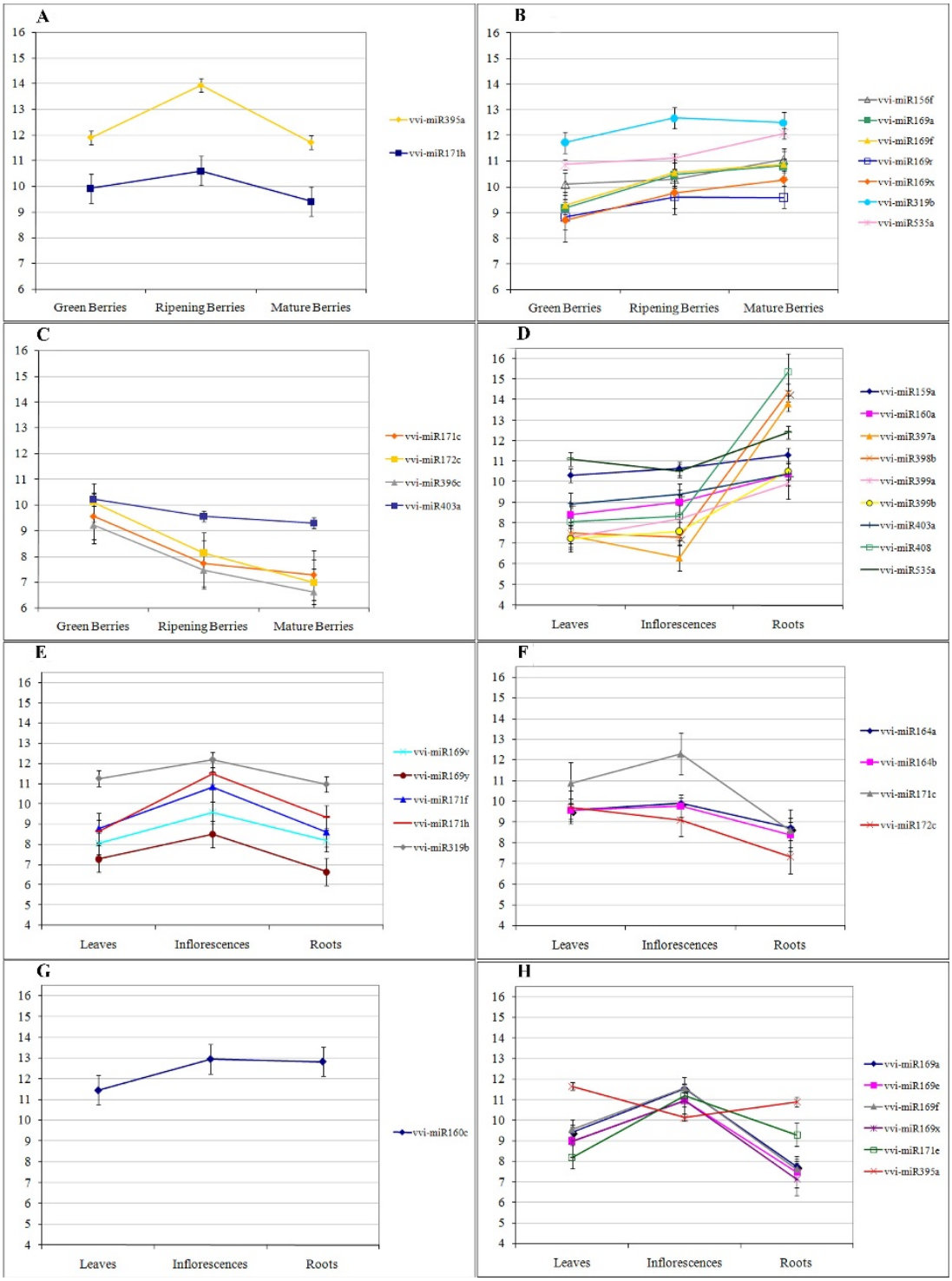


Fig 2.4 - -Differential expression of mature miRNAs by tissue. miRNAs showing significant changes in expression by tissue are reported. Panels A-C: miRNAs differentially expressed in one stage of berry ripening A: at veraison, B: in green berries, C: in mature berries. Panel D: miRNAs more highly expressed in roots, Panel E: miRNAs more highly expressed in inflorescences, Panel F: miRNAs less expressed in roots, Panel G: miRNAs less expressed in leaves, Panel H: miRNAs showing significant differences in all tissues tested. Error bars indicate confidence intervals. For all panels, the Y axis shows Log2 of the normalized median of spot intensities. From (Mica, Piccolo et al. 2010).

MiR396c shows 6 fold decrease in expression during ripening. The mir396 family targets seven *Growth Regulating Factor* (*GRF*) genes in *Arabidopsis* (Jones-Rhoades and Bartel 2004). *GRF* genes encode putative transcription factors associated with cell expansion in leaf and other tissues in *A. thaliana* and *O. sativa* (Kim, Choi et al. 2003; Choi, Kim et al. 2004). A potential role for miR396 in the regulation of cell expansion during fruit maturation is an intriguing hypothesis. In addition, recent data also link miR396 to responses to abiotic stresses including drought (Liu, Spizzo et al. 2008), again suggesting the importance of water homeostasis during berry ripening. miR172, downregulated during berry maturation, targets *Apetala 2* (AP2) -like transcription factors, regulators of flowering time, organ identity and of vegetative phase change (Lauter, Kampani et al. 2005). In grapevine, genes related to AP2 are upregulated at veraison, being involved in berry maturation (Terrier, Glissant et al. 2005) and putatively connected with abiotic and biotic stress resistance. This evidence fits well with our findings. The sharp up-regulation of miR395 at veraison suggests a further role for miRNAs in an agronomically important aspect of grape maturation. miR395 is known to contribute to the regulation of sulfur metabolism, targeting both sulfate transporters and ATP sulphurylase genes. A direct connection between ATP sulfurylases and berry maturation has not been demonstrated, but it is known that a Glutathione S-transferase is strongly connected with berry ripening and in particular with coloration during berry development (Terrier, Glissant et al. 2005).

MiR397a, miR398b and miR408 which are extremely highly expressed in root tissues target various copper proteins: plantacyanin, laccases and a superoxide dismutase, all putatively involved in stress responses and lignification (Jones-Rhoades and Bartel 2004; Sunkar and Zhu 2004; Lu, Sun et al. 2005; Sunkar, Kapoor et al. 2006). These miRNAs have also been shown to be coexpressed in *Arabidopsis* under conditions of copper deprivation (Abdel-Ghany and Pilon 2008). Moreover some *laccase* genes in *Arabidopsis* are root specific (for example *AtLAC15*) or mostly expressed in roots (McCaig, Meagher et al. 2005) and are involved in root elongation and lignification (Liang, Davis et al. 2006). Given that grapevine roots are much more lignified than those of *Arabidopsis*, it is plausible that regulation of laccase expression is vital in the grapevine. It is interesting to note that the laccase family is, along with other polyphenol oxidase gene families, massively expanded in grapevine with respect to *Arabidopsis* (>60 genes in *V. vinifera*, 17 in *Arabidopsis*).

Tab.2.1 – Oligoarray and deep sequencing results for miRNAs in grape

For each predicted pre-miRNA the table reports: **miRNA**=miRNA name; **Mature Sequence**=the mature sequence; **Leaf small**=the presence of perfect matching or oligoadenylated short RNA reads observed in leaf; **Illumina**= the presence of significant expression of the precursor observed by Illumina whole transcriptome sequencing; **454**= the presence of 454 reads including the precursor sequence in leaf; **Combimatrix**=Combimatrix oligoarray expression of mature sequence in at least one tissue. Mature miRNAs are ordered to reflect expected cases of crosshybridization for oligonucleotide arrays. For all microRNAs, chromosome (**Chr**) and strand (**Strand**), coordinates of the mature miRNA (**Start_M**=start, **Stop_M**=stop) and of the precursor miRNA (**Start_P**=start and **Stop_P**= stop) are provided into the 12× genome assembly of *Vitis vinifera*

At the end of the table are shown deep sequencing results for members of a new loci or family in 12x genome assembly (miRNA oligoarray was performed only on 8x genome assembly)

miRNA	Mature Sequence	Leaf small	Illumina	454	Combimatrix	Chr	Strand	Start_M	Stop_M	Start_P	Stop_P
VVI-MIR156A	TGACAGAAGAGAGGGAGCAC	no	no	no	yes	chr19	-	8708987	8709006	8708913	8709016
VVI-MIR156B	TGACAGAAGAGAGTGAGCAC	yes	yes	no	no	chr4	-	5357071	5357090	5356999	5357100
VVI-MIR156C	TGACAGAAGAGAGTGAGCAC	yes	no	no	no	chr4	-	848278	848297	848204	848307
VVI-MIR156D	TGACAGAAGAGAGTGAGCAC	yes	yes	no	no	chr11	-	7623290	7623309	7623217	7623319
VVI-MIR156E	TGACAGAGGAGAGTGAGCAC	no	no	no	no	chr11	-	1504272	1504291	1504194	1504301
VVI-MIR156F	TTGACAGAAGATAGAGAGCAC	yes	yes	no	yes	chr14	+	26463681	26463701	26463671	26463775
VVI-MIR156G	TTGACAGAAGATAGAGAGCAC	yes	yes	no	yes	chr17	-	3046396	3046416	3046324	3046426
VVI-MIR156I	TTGACAGAAGATAGAGAGCAC	yes	yes	no	yes	chr14	-	19727149	19727169	19727077	19727179
VVI-MIR156H	TGACAGAAGAGAGAGAGCAT	no	yes	yes	no	chr12	+	4108581	4108600	4108571	4108798
VVI-MIR159A	CTTGGAGTGAAGGGAGCTCTC	no	no	no	yes	chr15	-	18469183	18469203	18469173	18469366
VVI-MIR159B	CTTGGAGTGAAGGGAGCTCTC	no	no	no	yes	chr15	-	18471885	18471905	18471875	18472059
VVI-MIR159C	TTTGGATTGAAGGGAGCTCTA	yes	yes	no	yes	chr17	-	2609215	2609235	2609205	2609394
VVI-MIR160A	TGCCTGGCTCCCTGAATGCCAUC	yes	no	no	yes	chr12	-	10147229	10147249	10147157	10147259
VVI-MIR160B	TGCCTGGCTCCCTGAATGCCAUC	yes	no	yes	yes	chr10	+	1222389	1222409	1222379	1222482
VVI-MIR160C	TGCCTGGCTCCCTGTATGCCA	yes	yes	no	yes	chr10	+	11745671	11745691	11745661	11745766
VVI-MIR160D	TGCCTGGCTCCCTGTATGCCA	yes	no	no	yes	chr8	+	13017031	13017051	13017021	13017122
VVI-MIR160F	TGCCTGGCTCCCTGTATGCCA	yes	no	no	yes	chr13	+	5447926	5447946	5447916	5448017
VVI-MIR162	TCGATAAACCTCTGCATCCAG	yes	yes	yes	yes	chr17	+	4716591	4716611	4716519	4716621
VVI-MIR164A	TGGAGAAGCAGGGCACGTGCA	yes	no	no	yes	chr7	-	3287560	3287580	3287470	3287590
VVI-MIR164C	TGGAGAAGCAGGGCACGTGCA	yes	yes	no	yes	chr8	+	10080470	10080490	10080460	10080556
VVI-MIR164D	TGGAGAAGCAGGGCACGTGCA	yes	no	no	yes	chr14	-	1414652	1414672	1414565	1414682
VVI-MIR164B	TGGAGAAGCAGGGCACATGCT	no	no	no	yes	chr9	-	514820	514840	514758	514850
VVI-MIR166A	TCGGACCAGGCTTCATTCC	yes	yes	yes	yes	chr8	-	3302808	3302827	3302797	3302926
VVI-MIR166B	TCGGACCAGGCTTCATTCC	yes	yes	no	yes	chr12	+	17937465	17937484	17937398	17937496
VVI-MIR166C	TCGGACCAGGCTTCATTCCCC	yes	yes	no	yes	chr15	-	16978583	16978603	16978572	16978731
VVI-MIR166D	TCGGACCAGGCTTCATTCCCC	yes	yes	no	yes	chr16	-	21405227	21405247	21405216	21405375
VVI-MIR166E	TCGGACCAGGCTTCATTCCCC	yes	yes	no	yes	chr2	+	2255856	2255876	2255722	2255887

VVI-MIR166F	TCGGACCAGGCTTCATTCCCC	yes	no	no	yes	chr7	+	19450082	19450102	19450014	19450113
VVI-MIR166G	TCGGACCAGGCTTCATTCCCC	yes	no	no	yes	chr7	-	453869	453889	453858	453957
VVI-MIR166H	TCGGACCAGGCTTCATTCCCC	yes	yes	no	yes	chr5	-	6741199	6741219	6741188	6741287
VVI-MIR167A	TGAAGCTGCCAGCATGATCTG	no	yes	no	yes	chr1	+	1618514	1618534	1618504	1618854
VVI-MIR167B	TGAAGCTGCCAGCATGATCTA	yes	yes	yes	yes	chr14	+	7137398	7137418	7137388	7137486
VVI-MIR167C	TGAAGCTGCCAGCATGATCT	no	no	no	yes	chrUn	+	7495696	7495715	7495686	7495776
VVI-MIR167D	TGAAGCTGCCAGCATGATCTA	yes	yes	no	yes	chrUn	+	7490503	7490523	7490493	7490607
VVI-MIR167E	TGAAGCTGCCAGCATGATCTA	yes	no	no	yes	chr5	+	5845395	5845415	5845385	5845474
VVI-MIR168	TCGCTTGGTGCAGGTCGGGAA	yes	yes	yes	yes	chr2	-	17944902	17944922	17944801	17944932
VVI-MIR169B	TGAGCCAAGGATGGCTTGCCG	no	no	no	yes	chr11	+	16265447	16265467	16265437	16265541
VVI-MIR169H	TGAGCCAAGGATGGCTTGCCG	no	no	no	yes	chr11	+	16151661	16151681	16151651	16151751
VVI-MIR169A	CAGCCAAGGATGACTTGCCGG	no	no	no	yes	chr11	+	16112984	16113004	16112972	16113095
VVI-MIR169C	CAGCCAAGGATGACTTGCCGG	no	no	no	yes	chr4	-	2265982	2266002	2265925	2266014
VVI-MIR169J	CAGCCAAGGATGACTTGCCGG	no	no	no	yes	chr11	+	16101929	16101949	16101917	16102038
VVI-MIR169K	CAGCCAAGGATGACTTGCCGG	no	no	no	yes	chr11	+	16108551	16108571	16108539	16108660
VVI-MIR169S	CAGCCAAGGATGACTTGCCGG	no	no	no	yes	chr11	+	16439724	16439744	16439712	16439810
VVI-MIR169W	CAGCCAAGGATGACTTGCCGG	no	no	no	yes	chr14	+	29685626	29685646	29685614	29685756
VVI-MIR169L	GAGCCAAGGATGACTTGCCGT	no	no	no	yes	chr11	+	16185301	16185321	16185290	16185392
VVI-MIR169M	AGCCAAGGATGACTTGCCGG	no	no	no	yes	chr11	+	16361248	16361268	16361236	16361337
VVI-MIR169N	AGCCAAGGATGACTTGCCGGC	no	no	no	yes	chr11	+	16380252	16380272	16380240	16380340
VVI-MIR169O	GAGCCAAGGATGACTTGCCGC	no	no	no	yes	chr11	+	16190341	16190361	16190330	16190432
VVI-MIR169P	AGCCAAGGATGACTTGCCGGC	no	no	no	yes	chr11	+	16347822	16347842	16347810	16347911
VVI-MIR169Q	AGCCAAGGATGACTTGCCGGC	no	no	no	yes	chr11	+	16384496	16384516	16384484	16384580
VVI-MIR169E	TAGCCAAGGATGACTTGCCTGC	yes	yes	no	yes	chr14	-	25082834	25082854	25082717	25082865
VVI-MIR169F	CAGCCAAGGATGACTTGCCGA	no	yes	no	yes	chr1	+	12404220	12404240	12404208	12404391
VVI-MIR169G	CAGCCAAGGATGACTTGCCGA	no	yes	no	yes	chr8	+	21104463	21104483	21104451	21104571
VVI-MIR169R	TGAGTCAAGGATGACTTGCCG	no	no	no	yes	chr11	+	16415141	16415161	16415131	16415239
VVI-MIR169T	CGAGTCAAGGATGACTTGCCG	no	no	no	yes	chr11	+	16399577	16399597	16399567	16399676
VVI-MIR169U	TGAGTCAAGGATGACTTGCCG	no	no	no	yes	chr11	+	16409411	16409431	16409401	16409510
VVI-MIR169V	AAGCCAAGGATGAATTGCCGG	no	no	no	yes	chr11	+	16468036	16468056	16468025	16468120
VVI-MIR169X	TAGCCAAGGATGACTTGCCTA	no	no	yes	yes	chr17	-	355806	355826	355713	355837
VVI-MIR169Y	TAGCGAAGGATGACTTGCCTA	no	no	no	yes	chr1	+	22233583	22233603	22233573	22233820
VVI-MIR169I	GAGCCAAGGATGACTGGCCGT	no	no	no	yes	chr11	+	16158025	16158045	16158014	16158118
VVI-MIR169D	CAGCCAAGAATGATTTGCCGG	no	no	no	no	chr11	+	16106496	16106516	16106486	16106605
VVI-MIR171B	TGATTGAGCCGCGTCAATATC	no	no	no	yes	chr12	-	5542409	5542429	5542399	5542497
VVI-MIR171C	TGATTGAGCCGTGCCAATATC	yes	no	no	yes	chr12	-	5487760	5487780	5487747	5487849
VVI-MIR171D	TGATTGAGCCGTGCCAATATC	yes	no	no	yes	chrUn	-	40892395	40892415	40892382	40892481
VVI-MIR171A	TGATTGAGCCGTGCCAATATC	yes	yes	yes	yes	chr14	-	25491212	25491232	25491201	25491299
VVI-MIR171I	TGATTGAGCCGTGCCAATATC	yes	no	no	yes	chr17	+	893602	893622	893534	893636
VVI-MIR171E	TGATTGAGCCGCGCCAATATC	no	yes	no	yes	chr11	+	5203387	5203407	5203318	5203420

VVI-MIR171H	TGGTTGAGCCGCGCCAATATC	no	no	no	yes	chr17	-	1828663	1828683	1828653	1828748
VVI-MIR171F	TTGAGCCGCGCCAATATCACT	no	yes	no	yes	chr9	+	7012570	7012590	7012498	7012600
VVI-MIR171G	TTGAGCCGAACCAATATCACC	no	yes	no	yes	chr18	+	3255670	3255690	3255625	3255700
VVI-MIR172A	TGAGAATCTTGATGATGCTGCATC	yes	yes	no	no	chr6	-	17652420	17652440	17652412	17652521
VVI-MIR172B	TGAGAATCTTGATGATGCTGCATC	yes	no	no	no	chr13	-	6181378	6181398	6181370	6181487
VVI-MIR172D	AGAATCTTGATGATGCTGCAT	no	no	no	yes	chr8	+	12667251	12667271	12667173	12667281
VVI-MIR172C	GGAATCTTGATGATGCTGCAG	no	no	no	yes	chr13	+	3217614	3217634	3217507	3217644
VVI-MIR319B	TTGGACTGAAGGGAGCTCCCT	yes	no	no	yes	chr1	+	4189724	4189744	4189562	4189755
VVI-MIR319C	TTGGACTGAAGGGAGCTCCCT	yes	yes	no	yes	chr2	-	855572	855592	855561	855742
VVI-MIR319E	TTTGGACTGAAGGGAGCTCCT	no	yes	yes	yes	chr11	+	4317299	4317316	4317224	4317330
VVI-MIR319G	ATTGGACTGAAGGGAGCTCCC	yes	yes	no	yes	chr17	-	3675989	3676009	3675978	3676199
VVI-MIR319F	TTGGACTGAAGGGAGCTCCCT	yes	yes	no	yes	chr6	+	9137416	9137436	9137255	9137447
VVI-MIR390	AAGCTCAGGAGGGATAGCGCC	yes	yes	no	yes	chr6	+	8159529	8159549	8159519	8159658
VVI-MIR393A	TCCAAAGGGATCGCATTGATC	yes	no	yes	yes	chr16	-	17247283	17247303	17247187	17247312
VVI-MIR393B	TCCAAAGGGATCGCATTGATC	yes	yes	yes	yes	chr13	+	4265142	4265162	4265132	4265214
VVI-MIR394A	TTGGCATTCTGTCCACCTCCAT	yes	yes	no	no	chr12	-	17122063	17122082	17122004	17122092
VVI-MIR394B	TTGGCATTCTGTCCACCTCC	no	yes	no	no	chr18	-	1413101	1413120	1413038	1413130
VVI-MIR394C	TTGGCATTCTGTCCACCTCCT	yes	no	no	no	chr18	-	3551332	3551351	3551260	3551361
VVI-MIR395A	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	+	6527990	6528009	6527928	6528019
VVI-MIR395B	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	+	6502724	6502743	6502664	6502753
VVI-MIR395C	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	-	6499924	6499943	6499914	6500005
VVI-MIR395D	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	+	6512819	6512838	6512763	6512848
VVI-MIR395E	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	+	6505310	6505329	6505248	6505339
VVI-MIR395F	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	+	6489598	6489617	6489542	6489627
VVI-MIR395L	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	+	6559155	6559174	6559098	6559184
VVI-MIR395M	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	+	6557873	6557892	6557811	6557902
VVI-MIR395G	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	+	6482169	6482188	6482113	6482198
VVI-MIR395H	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	+	6566714	6566733	6566652	6566743
VVI-MIR395I	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	+	6562698	6562717	6562642	6562727
VVI-MIR395J	TGAAGTGTTTGGGGGAACTC	no	yes	no	yes	chr1	+	6553082	6553101	6553026	6553111
VVI-MIR395K	TGAAGTGTTTGGGGGAACTC	no	no	no	yes	chr1	+	6536841	6536860	6536780	6536871
VVI-MIR395N	CTGAAGAGTCTGGAGGAACTC	no	no	no	yes	chr17	-	6409005	6409025	6408995	6409131
VVI-MIR396B	TTCCACAGCTTTCTTGAACT	no	yes	no	yes	chr11	+	5246803	5246823	5246790	5246897
VVI-MIR396A	TTCCACAGCTTTCTTGAACTA	no	yes	no	yes	chr9	-	7372606	7372624	7372522	7372637
VVI-MIR396C	TTCCACAGCTTTCTTGAACTG	no	no	no	yes	chr4	-	5119670	5119688	5119591	5119698
VVI-MIR396D	TTCCACAGCTTTCTTGAACTG	no	yes	no	yes	chr11	-	5253201	5253219	5253110	5253229
VVI-MIR397A	TCATTGAGTGCAGCGTTGATG	yes	yes	no	yes	chrUn	-	11971983	11972003	11971898	11972015
VVI-MIR398A	TGTGTTCTCAGGTCACCCCTT	yes	yes	no	yes	chr1	+	731684	731704	731608	731710
VVI-MIR398B	TGTGTTCTCAGGTCGCCCCTG	yes	yes	no	yes	chr6	-	16503558	16503578	16503544	16503631
VVI-MIR398C	TGTGTTCTCAGGTCGCCCCTG	yes	yes	no	yes	chr6	+	15575634	15575654	15575581	15575668

VVI-MIR399A	TGCCAAAGGAGAATTGCCCTG	no	yes	no	yes	chr10	+	2989516	2989536	2989450	2989546
VVI-MIR399H	TGCCAAAGGAGAATTGCCCTG	no	yes	no	yes	chr10	+	2983604	2983624	2983545	2983634
VVI-MIR399B	TGCCAAAGGAGAGTTGCCCTG	no	no	no	yes	chr16	-	15618718	15618738	15618708	15618824
VVI-MIR399C	TGCCAAAGGAGAGTTGCCCTG	no	no	no	yes	chr15	+	15232251	15232271	15232200	15232281
VVI-MIR399I	CGCCAAAGGAGAGTTGCCCTG	yes	yes	no	yes	chr2	+	4101895	4101912	4101801	4101922
VVI-MIR399D	TGCCAAAGGAGATTTGCTCGT	no	no	no	no	chr10	-	2988031	2988051	2988021	2988125
VVI-MIR399E	TGCCAAAGGAGATTTGCCCGG	no	yes	no	no	chr10	-	2992230	2992250	2992220	2992331
VVI-MIR399F	TGCCGAAGGAGATTTGTCCTG	no	no	no	no	chr10	-	2995818	2995838	2995807	2995913
VVI-MIR399G	TGCCAAAGGAGATTTGCCCCT	no	no	no	yes	chr10	-	2981257	2981277	2981247	2981359
VVI-MIR403A	TTAGATTCACGCACAAACTCG	yes	no	no	yes	chr5	+	65331	65351	65247	65361
VVI-MIR403B	TTAGATTCACGCACAAACTCG	yes	no	no	yes	chr5	+	600236	600256	600181	600266
VVI-MIR403C	TTAGATTCACGCACAAACTCG	yes	yes	no	yes	chr5	+	602695	602715	602611	602725
VVI-MIR403D	TTAGATTCACGCACAAACTCG	yes	no	no	yes	chr5	+	166537	166557	166482	166567
VVI-MIR403E	TTAGATTCACGCACAAACTCG	yes	no	no	yes	chr5	+	168183	168203	168099	168213
VVI-MIR403F	TTAGATTCACGCACAAACTCG	yes	yes	no	yes	chr7	-	4179683	4179703	4179673	4179780
VVI-MIR408	ATGCACTGCCTCTTCCCTGGC	yes	yes	yes	yes	chr7	+	5012012	5012031	5011935	5012041
VVI-MIR477	ATCTCCCTCAAAGGCTTCCAA	no	no	no	yes	chr1	+	22740271	22740291	22740261	22740354
VVI-MIR479	TGTGGTATTGGTTCGGCTCATC	yes	no	no	no	chr16	+	21573769	21573790	21573759	21573852
VVI-MIR482	TCTTTCCTACTCCTCCCATTCC	yes	yes	yes	no	chr17_random	+	5523114	5523130	5523024	5523140
VVI-MIR535A	TGACAACGAGAGAGAGCACGC	yes	no	yes	yes	chr7_random	-	1392322	1392343	1392253	1392353
VVI-MIR535B	TGACAACGAGAGAGAGCACGC	yes	no	yes	yes	chrUn	-	25369772	25369793	25369703	25369803
VVI-MIR535D	TGACAACGAGAGAGAGCACGC	yes	no	yes	yes	chr7_random	-	1346437	1346458	1346368	1346468
VVI-MIR828A	TCTTGCTCAAATGAGTATTCCA	no	no	no	no	chr16	+	21724317	21724338	21724308	21724429
VVI-MIR828B	TCTTGCTCAAATGAGTGTTCCA	no	no	no	no	chr1	+	2961655	2961676	2961645	2961716
VVI-MIR845A	TAGCTCTGATACCAATTGATA	no	no	no	no	chr16	+	20505160	20505180	20504749	20505190
VVI-MIR845B	TAGCTCTGATACCAATTGATA	no	no	no	no	chr2	+	12265254	12265274	12264843	12265284
VVI-MIR845C	AGGCTCTGATACCAATTGATG	no	no	no	no	chr7	-	207788	207808	207778	207867
VVI-MIR845D	TGGCTCTGATACCAATTGATG	no	no	no	no	chr4	-	9870414	9870434	9870404	9870534
VVI-MIR845E	TGGCTCTGATACCAATTGATG	no	no	no	no	chr5	+	13374777	13374797	13374677	13374807
VVI-MIR171J_new_locus	TGATTGAGCCGTGCCAATATC	yes	no	no	nd	chr18	+	1502398	1502418	1502310	1502425
VVI-MIR159D_new_locus	TTTGGTTTGAAGGGAGCTCTG	no	no	no	nd	chr15	-	18466307	18466327	18466309	18466475
VVI-MIR166I_new_locus	TCGGACCAGGCTTCATTCCCC	no	no	no	nd	chr7	+	19450307	19450327	19450174	19450334
VVI-MIR396E_new_locus	TTCCACGGCTTTCTTGAACTT	yes	yes	no	nd	chr1	+	1997823	1997843	1997815	1997984
VVI-MIR399J_new_locus	TGCCAAAGGAGATTTGCCCCG	no	no	no	nd	chr10	-	2978909	2978929	2978909	2979003
VVI-MIR172E_new_locus	GGAATCTTGATGATGCTGCAT	no	no	no	nd	chr6	+	6003709	6003729	6003613	6003731
VVI-MIR395O/P_new_locus	TGAAGTGTTTGGGGGAACTC	no	no	no	nd	chr1	+	6517117	6517136	6516957	6517141
VVI-MIR390B_new_locus	AAGCTCAGGAGGGATAGCGCC	yes	no	no	nd	chr8	+	9571422	9571442	9571421	9571519
VVI-MIR530_new_family	TGCATTTGCACCTGCACCTT	yes	no	no	nd	chr8	-	18999858	18999877	18999728	18999886
VVI-MIR827_new_family	TTAGATGATCATCAACAAACA	yes	yes	no	nd	chr5	-	24742138	24742158	24742131	24742224

2.4 MiRNA expression and deep sequencing data

2.4.1 The smallRNA deep sequencing approach

The first and still most common approach to the discovery of novel small RNAs has been to clone and sequence individual small RNAs using traditional molecular methods. The majority of known miRNAs were identified by this approach. In fact it was first used to identify miRNAs and siRNAs in mammals, *Caenorhabditis elegans, Drosophila* and *Arabidopsis* (Lagos-Quintana, Rauhut et al. 2001; Lau, Lim et al. 2001; Lee and Ambros 2001; Llave, Kasschau et al. 2002; Park, Li et al. 2002).

This approach is based on the property that during microRNAs biogenesis both strands of microRNA precursors are processed by RNase III into small RNA segments (20~24 nt) that have 5′ phosphateand 3′ hydroxyl termini, in contrast to most RNA turnover products that have a 5′ hydroxyl terminus (Zamore, Tuschl et al. 2000). For cloning are used protocols that require the presence of 5′phosphateand free 3′ hydroxyl group on the small RNAs for adapter ligation. In particular, after reverse transcription, the cDNA is PCR-amplified using primers corresponding to the adapter sequences. PCR products are cloned and then sequenced. It is known that about 30–50% of the clones represent RNA turnover products of the abundant rRNAs, tRNAs, snRNAs (Llave, Kasschau et al. 2002; Sunkar, Girke et al. 2005). The cloning frequency of an individual small RNA generally reflects its relative abundance in the sample, providing a quantitative expression measurement.

Despite the early success of this approach, it is unlikely that these efforts are saturating for rare or tissue-specific small RNAs.

The recent introduction of deep sequencing technology, enabling the simultaneous sequencing of up to millions of DNA or RNA molecules, has provided another option for profiling microRNAs (Creighton, Reid et al. 2009).

Compared to microarrays, deep sequencing technology for profiling mRNA expression remains rather expensive. Currently it is trying to meet the goal of generating complete human genome sequences for less than $100 000 (Schloss 2008). However, deep sequencing overcomes many of the disadvantages of microarrays, which suffer from background and cross-hybridization problems and measure only the relative abundances of previously discovered microRNAs. In addiction, profiling the small RNA fraction that contains microRNAs is much more feasible; deep sequencing measures absolute abundance and is not

limited by array content. In this way allows the discovery of novel microRNAs or, maybe other small RNA species (Creighton, Reid et al. 2009).

The detection and quantification of small RNAs using deep sequencing methods was first attained in *Arabidopsis* by Meyers lab (Lu, Meyers et al. 2007). More than 2,000,000 small RNAs were sequenced by Massively Parallel Signature Sequencing (MPSS) (Brenner, Johnson et al. 2000) from *Arabidopsis* flowers and seedlings, yielding more than 70,000 genome-matching distinct sequences. This symbolizes a significant advance over more traditional methods for small RNA identification. One of MPSS limitations is that it is only able of sequencing the 5′ 17 nucleotides of small RNAs.

Solexa, Inc. has developed a four-color DNA sequencing-by-synthesis (SBS) method as a replacement for MPSS based on a novel, reversible, dye-termination chemistry (http://www.solexa.com). This approach can generate more than 10,000,000 tags (5 times more than MPSS approach!) with high accuracy.

Another approach, the Supported Oligo Ligation Detection or SOLiD, uses an array of microbeads each coated with a single DNA or cDNA fragment; a pool of fluorescent oligos is used to "read" the sequences by complementary binding using a repeated process of ligation, detection, and cleavage. This determines up to 50 nucleotides of sequence for bead, for >10 million beads.

These novel, highly parallel methods have the potential to dramatically reduce the cost of sequencing and offer a much richer source of sequence information.

2.4.2 Illumina Sequencing

The Illumina Genome Analyzer sequencing system is a ground breaking platform for genetic analysis and functional genomics. It dramatically improves speed and reduces costs, transforming the way many experiments are devised and carried out.

Step 1: Sample Preparation

For smallRNA sequencing, total RNA is ligated to specific 3' and 5' adapter sequences and reverse transcribed. Size selection is performed to isolate ligation products that should contain inserts in the 20-30 base size range.

Steps 2-6: Cluster Generation by Bridge Amplification

Illumina utilizes a unique "bridged" amplification reaction that occurs on the surface of the flow cell.

The flow cell surface is coated with single stranded oligonucleotides that correspond to the sequences of the adapters ligated during the sample preparation stage. Single-stranded, adapter-ligated fragments are bound to the surface of the flow cell exposed to reagents for polyermase-based extension. Priming occurs as the free/distal end of a ligated fragment "bridges" to a complementary oligo on the surface (Bentley, Balasubramanian et al. 2008).

Repeated denaturation and extension results in localized amplification of single molecules in millions of unique locations across the flow cell surface. This process occurs in what is referred to as Illumina's "cluster station", an automated flow cell processor.

Steps 7-12: Sequencing-by-synthesis (SBS)

A flow cell containing millions of unique clusters is now loaded into the sequencer for automated cycles of extension and imaging.

The first cycle of sequencing consists first of the incorporation of a single fluorescent nucleotide, followed by high resolution imaging of the entire flow cell. These images represent the data collected for the first base. Any signal above background identifies the physical location of a cluster (or polony), and the fluorescent emission identifies which of the four bases was incorporated at that position (Bentley, Balasubramanian et al. 2008; Morozova and Marra 2008).

This cycle is repeated, one base at a time, generating a series of images each representing a single base extension at a specific cluster. Base calls are derived with an algorithm that identifies the emission colour over time. At this time reports of useful Illumina reads range from 26-100 bases.

The use of physical location to identify unique reads is a critical concept for all next generation sequencing systems. The density of the reads and the ability to image them without interfering noise is vital to the throughput of a given instrument. Each platform has its own unique issues that determine this number, 454 is limited by the number of wells in their PicoTiterPlate, Illumina is limited by fragment length that can effectively "bridge", and all providers are limited by flow cell real estate.

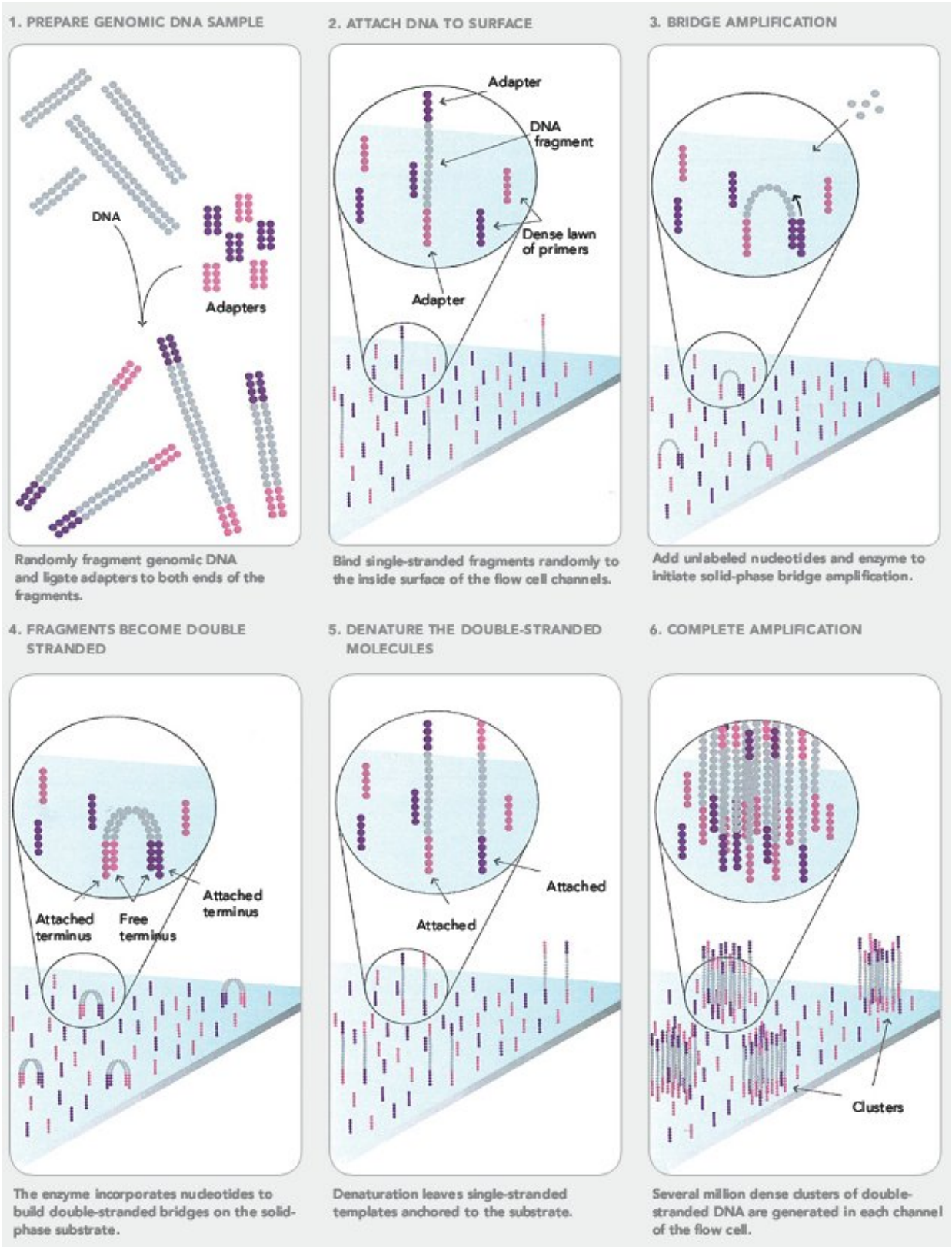


Fig.2.5 - The Illumina Genome Analyzer sequencing system
Sample Preparatio (step 1) and Cluster Generation by Bridge Amplification (steps 2-6)

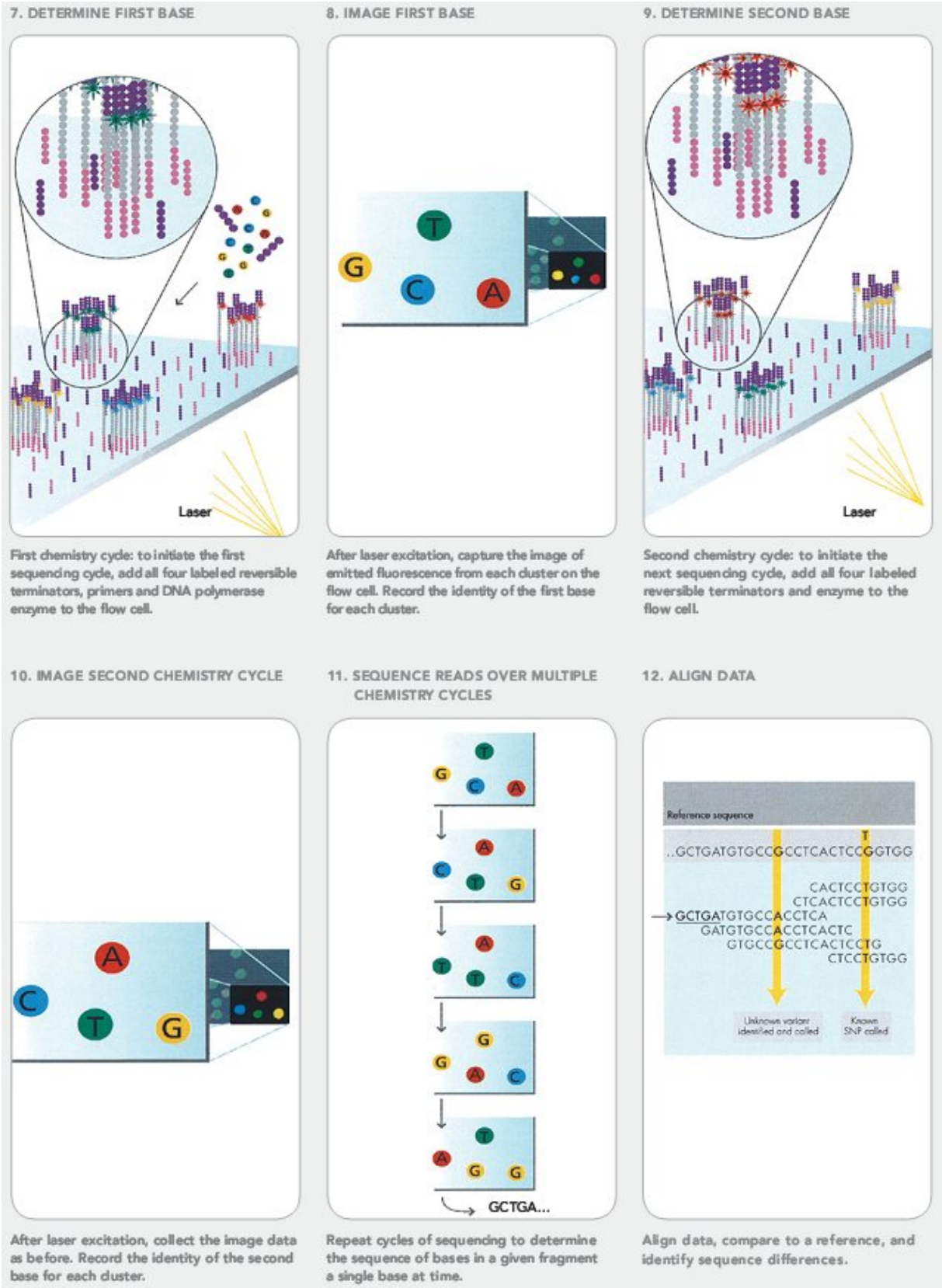


Fig.2.6 - The Illumina Genome Analyzer sequencing system
Sequencing-by-synthesis (SBS) (steps 7-12)

2.4.3 Results and discussion - *Deep sequencing of small RNAs from grapevine leaf tissue*

We generated 13,078,222 reads with Illumina sequencing of small RNA isolated from *Vitis vinifera* L. clone PN40024 leaves.

The Illumina sequence reads are 35 bases long, although the insert size range of interest is from 19-28 bases. Accordingly, sequence reads are expected to contain the first few bases of the 3' adapter sequence used in sample preparation (Fig.2.7). We used a custom script to identify and remove adapter sequences from the sequence reads. This script allows up to 2 mismatches in the adapter sequence to accommodate erroneous base calls in the extreme 3' end of the reads.

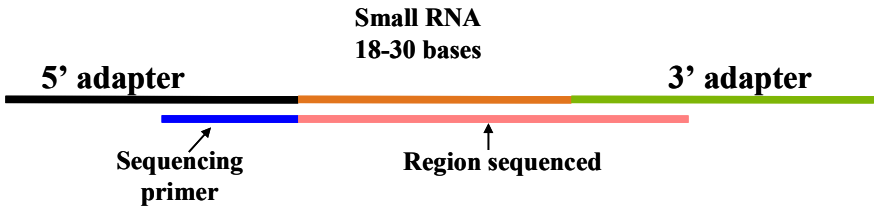


Fig.2.7 - Removal of adapter sequences from the sequence reads
Reads are typically at least 35 bases in length, but the small RNAs are 18- 30 bases. We must trim the 5' end of the 3' adapter . Sequence errors tend to be more frequent in the 3' part of the read, so it is a good idea to use a method that will accept at least one mismatch with the adapter sequence (potentially considering quality scores)

2,585,821 individual small RNA reads of 18-27 bases (19.8% of the total reads generated) yielded at least one perfect match to the draft genome after removal of adapter sequences and allowing for post transcriptional oligoadenylation of reads.

We mapped tags with the software SOAP (short oligonucleotide alignment) because respect to other short oligonucleotide alignment programs the comparison between the performance (time consumed) and sensitivity (reads aligned) is the best (Li, Li et al. 2008).

After exclusion of reads mapping on annotated structural RNAs, over 7% of the total mapped sequences were of length 21 bases and accounted for 7.8% of the genomic loci represented by the mapped data (mean redundancy of 4.38 reads/locus).

15% of loci represented were of length 24 (10.7% of tags sequenced) with a mean redundancy of 3.08 reads/locus, suggesting, in accord with other studies (Vaucheret 2006; Moxon, Jing et al. 2008), that miRNAs in our sample tend to be expressed at higher levels or processed more specifically than the more heterogeneous 24 base small RNAs.

After mapping of reads to the *Vitis* genome, custom scripts were used to count which proposed mature miRNA loci produced perfectly matching Illumina smallRNA reads.
In table Tab.2.1, the column entitled "Leaf small" shows the results of this analysis. Mapping of the short tags onto the genome sequence revealed that of the 30 families predicted by our comparative analysis, 26 showed at least one sequence tag either in exact or very close correspondence to the position of one of the predicted mature sequences (the exceptions being miR395, miR477, miR828 and miR845)

2.5 Whole transcriptome sequencing and differential expression of precursors

In plants, the majority of pri-miRNA transcripts are polyadenylated as they are transcribed by RNA polymerase II (Xie, Lu et al. 2005). However, the physiological half life of primary miRNA transcripts is expected to be short. Notwithstanding this limitation, we hypothesize that sequences derived from highly expressed pri-miRNA transcripts should be represented in whole transcriptome "deep sequencing" experiments as well as, potentially, in EST collections.
We have analyzed whole polyA$^+$ transcriptome data generated by the French Italian public Consortium for Grapevine Genome Characterization with the Illumina Solexa technology (Denoeud, Aury et al. 2008) and Roche 454 next generation sequencing platforms (unpublished data).

2.5.1 Illumina Solexa technology: polyA$^+$ RNA

Tanscriptome data, prepared with a similar strategy to the leaf smallRNA data, but commencing with polyA+ RNA and excluding the size fractionation step was generated from *in vitro* cultivated *Vitis vinifera* pn40024 plants (juvenile leaf, juvenile stem, juvenile root and embryonic callus) (Denoeud, Aury et al. 2008).

A total of 135,047,735 Illumina sequences (33-35 bases in length) derived from polyA$^+$ RNA isolated from 4 tissues were considered. The number of sequences detected for each tissues are in Tab.2.2.
Tags were mapped to the grapevine genome using the SOAP software (Li, Li et al. 2008) and coordinates were compared to those for predicted pre-miRNAs by microHARVESTER software (Dezulian, Remmert et al. 2006). For a limited number of precursor loci, as we expected, tags are spread more or less all the precursor length.

The statistical significance of the number of reads mapping within a predicted pre-miRNA was evaluated.

We estimate the probability that at least the observed number of reads should be clustered in the genomic interval defined by the precursor using the Poisson distribution. Thus, we exclude all reads mapping to predicted genes, and search for significant violations (at the 1% confidence interval) of the null-hypothesis that remaining reads should be distributed randomly among intergenic regions. We consider only reads mapping uniquely to a single genomic locus. Given the expected short half-life of most primary miRNA transcripts, we believe that these criteria constitute an extremely conservative test of precursor expression.

Tissue	Number of detected sequences
juvenile leaf	29,829,113
juvenile stem	30,785,175
juvenile root	29,254,635
embryonic callus	45,178,812

Tab.2.2 - Number of detected sequence for each tissue.
From a total of 135,047,735 Illumina sequences (33-35 bases in length) derived from polyA$^+$ RNA isolated from juvenile leaf, juvenile stem, juvenile root and embryonic callus.
RNA-Seq data are available from http://www.genoscope.cns.fr/externe/gmorse/raw_data/.

52 predicted precursors show significant expression in at least one tissue (25 in leaf, 38 in stem, 17 in root, 33 in callus).

The column correspondent to the attribute "Illumina" of Tab.2.1 show the results of polyA$^+$ transcriptome data analysis with Illumina transcriptome data are summarized. The entry "yes" signifies significant levels of expression in at least one tissue.

Many predicted precursors show a wide expression (miR156d, miR159c, miR166a and c, miR168, miR171a, miR398a, miR398b and c, miR408, miR482). In some families, when expressed, precursors show overlapping patterns. For example, miR319c, miR319e and miR319f are all expressed in stem, while miR319c and miR319g are expressed in callus, no expression of miR319 was detected in leaf or root. A similar situation is observed for the

miR396 family. In other cases, different precursors seem to be predominantly expressed in different tissues. For example miR171e transcripts are detected only in callus, miR171f is only transcribed in stem while miR171g is observed in callus and root - a similar situation can be observed for several families including miR166, miR167 and miR169). These data suggest that tissue specific expression of different precursors within single families is widespread in the grapevine.

Other precursors seem to be predominantly transcribed in specific tissues. For example miR171e transcripts are detected only in callus, miR171f is only transcribed in stem while miR171g is observed in callus and root. A similar situation can be observed for several families including miR166, miR167 and miR169.

2.5.2 454 transcriptome analysis

For 454 transcriptome analysis, polyA$^+$ RNA was isolated from *V. vinifera* L. cv Corvina leaf and berry tissues(Rezaian and Krake 1987).

454 deep sequencing analysis generated 613,098 reads from leaf and 581,655 from berry. These reads were mapped to the *Vitis genome* using the software BLAT (Kent 2002). Custom scripts were used to collect preliminary coordinates that coincided with harvester predictions, SPIDEY (Wheelan, Church et al. 2001) was used for fine mapping of splice junctions.

The column correspondent to the attribute “454” in Tab.2.1 shows for which predicted miRNA loci expression was supported by 454 transcriptome sequencing data. Expression of 15 loci received further support from these data. With the exception of miR160b and the miR535 family the expression of all precursors detected by 454 sequencing in leaf was also strongly supported by Illumina data.

2.5.3 Results - Estimation of primary microRNA transcripts and splice sites

For a number of predicted microRNAs the density of coverage of the corresponding genomic loci was sufficient to attempt to estimate patterns of splicing and alternative splicing.

PolyA$^+$ RNAs are usually mature messages and don’t contains introns. As consequence of this, it is very hard to map into the genome tags located in correspondence of the junction of 2 exons (Fig.2.8). For this reason, sometimes in case of splicing we have a discontiguous mapping of whole transcriptome reads.

In order to solve this problem, we developed a strategy that exploit the property that most introns have a GT at the 5'-splice site (donor site) and an AG at the 3' splice site (acceptor site).

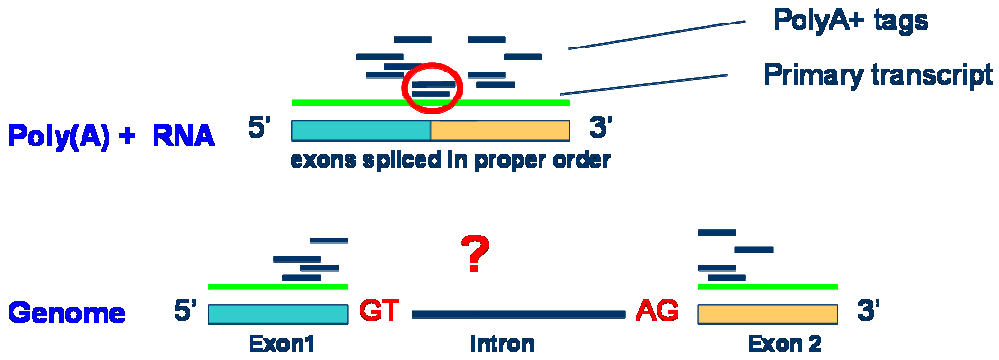


Fig.2.8 - Problem mapping: tags of spliced regions
Poly(A)+RNA tags doesn't contain introns and tags falling into spliced regions of genomes are not mapped by SOAP program.
GT=donor site; AG =acceptor site.

We scanned grape genome regions 5 kb upstream and downstream of predicted miRNA precursors for all GT dinucleotides (to define candidate donor sites) and all the AG dinucleotides (to define candidate acceptor sites). We collected all genome sequences mapping before the donor and then concatenated each of them to sequences mapping after the acceptor site. Finally, we mapped reads that did not provide perfect matches to the genomic sequence onto these conceptually spliced sequences. We identified all matches where reads provided perfect matches with at least 8 bases on either side of the splice junction. Introns inferred from mapping of 454 transcriptome reads were also recorded.

Fig.2.9 shows an example of the application of our strategy: We found 5 different possible canonical splice junctions for a specific locus, but only one of those corresponds to a real splicing product (Fig.2.9). We defined the transcription profile of vi-miR394b precursor in each tissue and detect the presence of a canonical intron supported by 14 Illumina reads (7 distinct sequences). This intron was also easily detectable through RACE experiments (performed by Dott.ssa Erica Mica). We note that the position of the intron corresponds well to a region of low, or undetectable levels of Illumina transcriptome coverage. Vvi-miR394B appears to be transcribed in callus, stem and leaf, while is not in root.

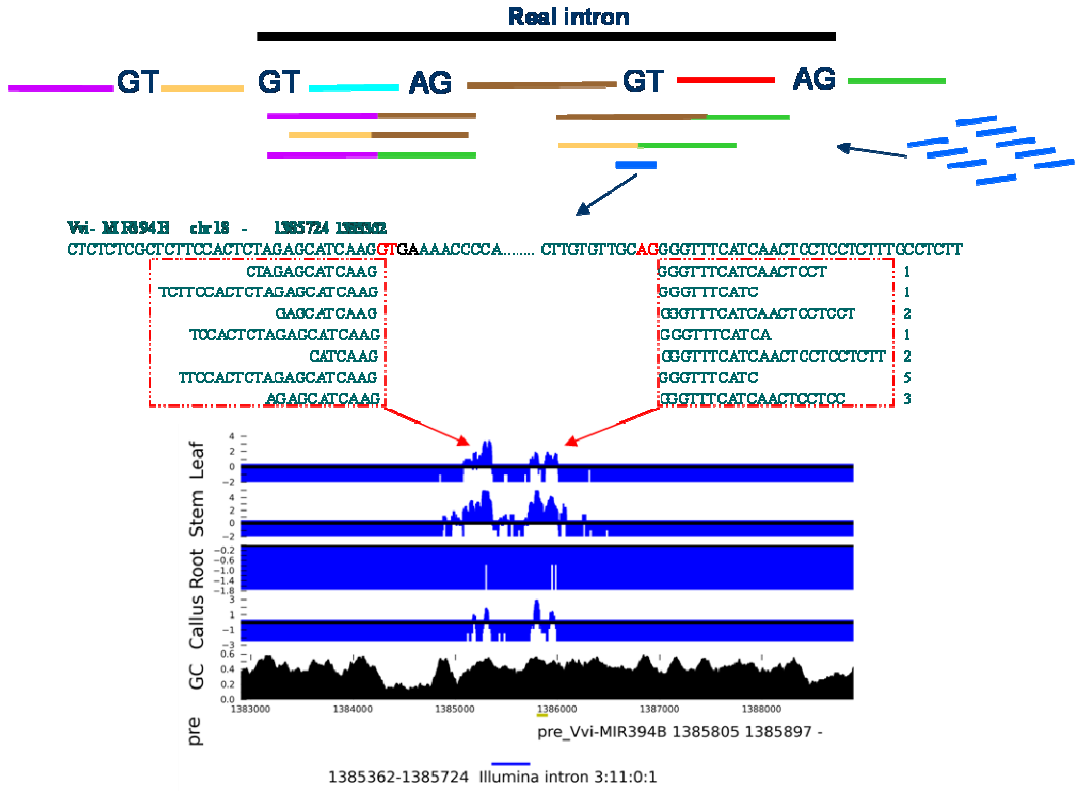


Fig.2.9 - Transcription and splicing of pri-miR394b in *Vitis vinifera*.
A summary of transcription of genomic loci containing predicted pre-miRNAs is provided. Illumina whole transcriptome reads per base are reported for four tissues as log(number of reads/expected number of reads under random distribution of reads). Local GC content, position and strand of predicted pre-miRNA as also shown along with coordinates of: canonical introns inferred from non-contiguous mapping of Illumina reads (blue bars), 454 reads (black bars) and assembled 454 sequence contigs (green bars). Predicted genes where present are represented by red bars.
GT donor and AC acceptor are coloured in red and defines the intron boundaries.
Sequences of tags (mapping in correspondence of the junction of 2 exons) are outlined by the dashed red. From (Mica, Piccolo et al. 2010)modified.

In the case of vvi-MIR162 we found evidence of alternative splicing isoforms (confirmed by 454 transcriptome data) (Fig.2.10).

Indeed, while the boundaries of proposed introns correspond to "shoulders" of falling transcript coverage, significant levels of reads mapping within the putative intronic sequences are observed. This observation is also consistent with the occurrence of alternative splicing and may indicate the presence of other, non detected alternative splicing events.

Interestingly, Hirsch et al. (Hirsch, Lefort et al. 2006) demonstrated that the primary miR162a transcript of *Arabidopsis* is subjected to complex pattern of alternative splicing, similar to that proposed for the grapevine miR162 transcript. In *Arabidopsis* are present at least four transcript isoforms (Fig.2.11), but only one of these (the unspliced one) leads to the

correct miRNA hairpin. In particular the npcRNA 78 gene contains the miR162 sequence in an alternative intron and corresponds to the MIR162a locus (Hirsch, Lefort et al. 2006).

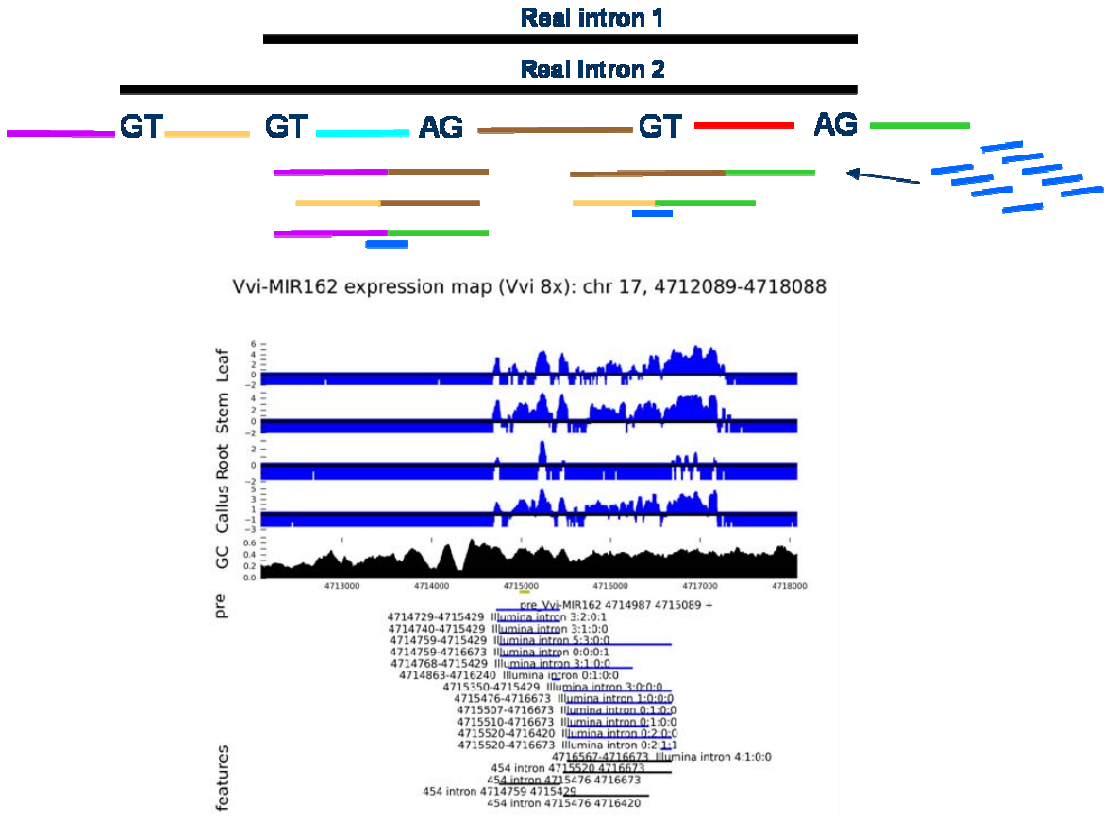


Fig.2.10 - Transcription and alternative splicing of pri-miR162 in *Vitis vinifera*.
Whole transcriptome reads per base are reported for four tissues as log(number of reads/expected number of reads under random distribution of reads). Local GC content, position and strand of predicted pre-miRNA as also shown along with coordinates of: canonical introns inferred from non-contiguous mapping of Illumina reads (blue bars), 454 reads (black bars) and assembled 454 sequence contigs (green bars). Predicted genes where present are represented by red bars. From (Mica, Piccolo et al. 2010) modified.

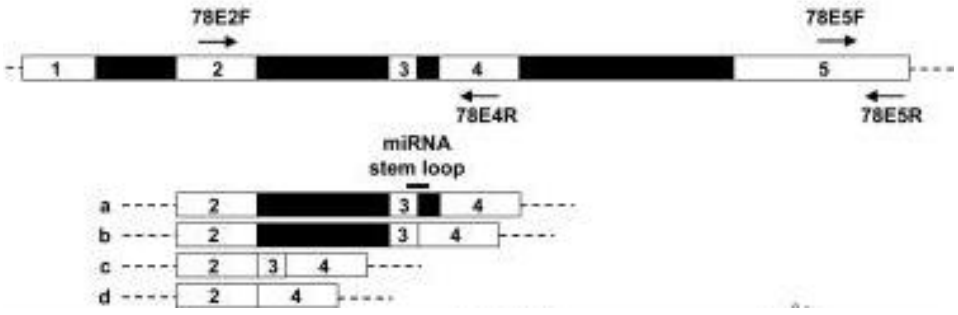


Fig.2.11 - Diagrammatic representation of the differentially spliced transcripts of npcRNA 78
With 'a','b','c','d' are shown the different isoforms related to npcRNA 78. With black box are indicated introns, with white box numbered are indicated exons. The isoform that corresponds to the functional pre-miRNA is the 'a', in which the intron between the exon 3 and 4 is retained. From (Hirsch, Lefort et al. 2006) modified

Interestingly, miR162 is a negative regulator: it inactivates DCL1 (DICER-like1) that contributes to the nuclear processing of all miRNAs (miR162 included). In other words, DCL1 generates active miR162 that blocks DCL1 production creating a negative feedback loop (Xie, Kasschau et al. 2003). The ulterior control at the transcription level by splicing underlines the importance miR162 as regulator.

In fact mir162 works in a complex mechanism for regulating the development surface of leaf. In the absence of miRNA162 , DCL acts on miR165/166 precursor generating the active miRNAs. These active miRNAs in turn block translation of two transcription factor genes, PHV/PHB. When no miR165/166 is present in the cytosol of the leaf primordial cells, the upper surface (adaxial) is developed. On the other hand, their presence turns off the two transcription factor genes, leading to the development of the lower (abaxial) surface of the leaf. miR165/166 is normally found at positions distant from the meristem (Carrington and Ambros 2003; Gustafson, Allen et al. 2005).

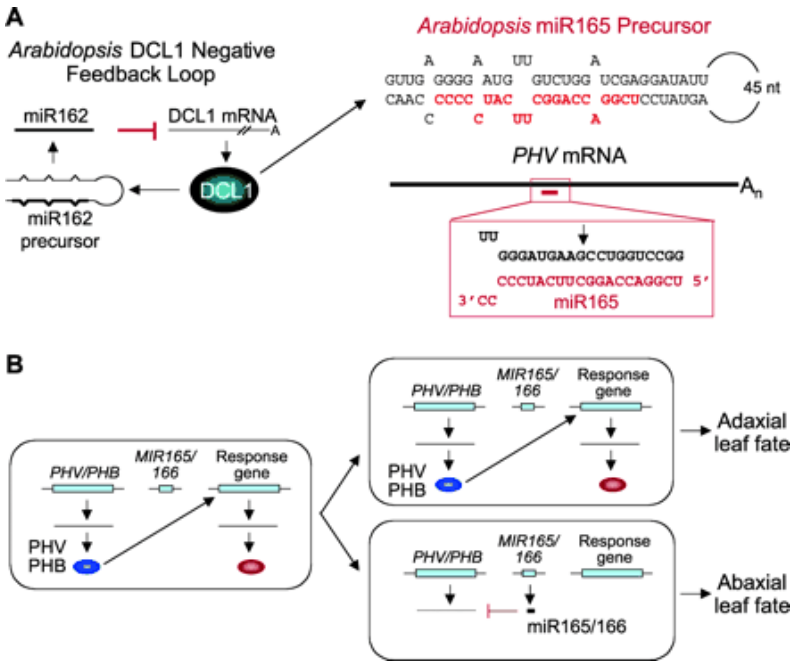


Fig.2.11 - Structure and function of *Arabidopsis* miRNAs.
(**A**) Expression of *DCL1*, which catalyzes miRNA precursor processing, is under negative-feedback regulation by miR162 (left). miR165/166 negatively regulates *PHV* and *PHB* mRNAs by guiding sequence-specific cleavage (right). *PHV*and *PHB* are related genes encoding HD-Zip transcription factors. miR165 and miR166 are related miRNAs that are predicted to interact with *PHV*and *PHB* mRNAs. Only *PHV* mRNA and miR165 are represented. Arrow, miR165-guided cleavage site.
(**B**) Model for specification of adaxial/abaxial polarity in *Arabidopsis* leaves. Expression of *PHV* and *PH* in leaf primordium cells close to the meristem results in a transcription program specifying adaxial fate. Inhibition of *PH* and *PHB* by miR165/166-guided degradation in cells distant to the meristem specifies abaxial fate.
From (Carrington and Ambros 2003)

Our findings suggest conservation of alternative splicing as a key regulatory mechanism in miR162 expression and indicate that Illumina and 454 transcript data can also be used to identify alternatively spliced plant pri-miRNAs.

Fig.2.12 shows evidence for expression of the miR168 locus. Analogously to miR162, our data suggest alternative splicing of the pri-mRNA, while the distribution of 454 contigs is highly consistent with the Illumina data.

Vaucheret et al. (Vaucheret 2006) showed that AGO1, the target of miR168 is involved in the regulation of miR168 stability. Our data may hint at yet another mechanism of regulation of this intriguing miRNA.

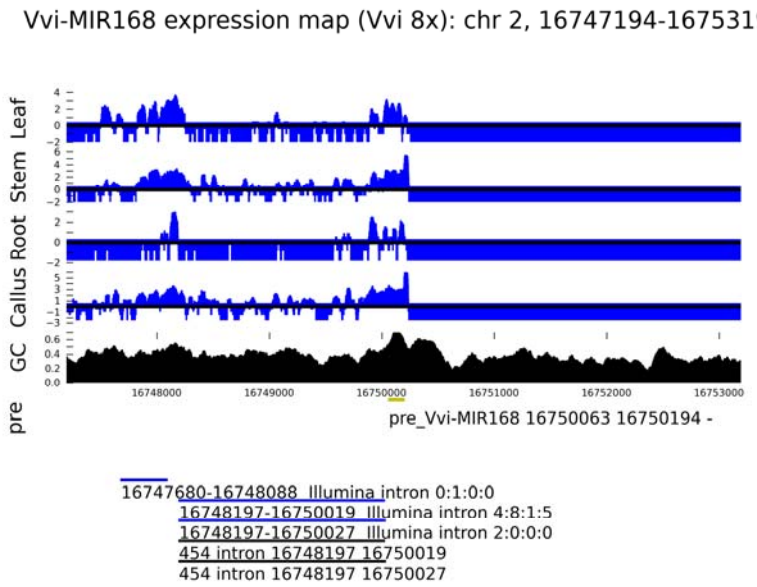


Fig.2.12 - Transcription and alternative splicing of pri-miR168 in *Vitis vinifera*.
Whole transcriptome reads per base are reported for four tissues as log(number of reads/expected number of reads under random distribution of reads). Local GC content, position and strand of predicted pre-miRNA as also shown along with coordinates of: canonical introns inferred from non-contiguous mapping of Illumina reads (blue bars), 454 reads (black bars) and assembled 454 sequence contigs (green bars). Predicted genes where present are represented by red bars. From (Mica, Piccolo et al. 2010) modified

Of 25 precursor loci chosen on the basis of extensive RNA-Seq, 18 showed evidence of transcript splicing and 8 of alternative splicing, suggesting that post-transcriptional modification of miRNA transcripts is likely to be widespread.

In Fig. 2.13 is shown the junction read coverage for vvi-miR394b, vvi-miR162 and vvi-miR168.

Fig.2.13 - Splice junction read coverage for vvi-miR394b, vvi-miR162 and vvi-miR168
Donor site (GT) and acceptor site /AG) are in red.

It is possible that some splicing events frequently identified by deep sequencing approaches could be associated with regulation of downstream processing of transcripts as for the miR162 transcript of *Arabidopsis* (Hirsch, Lefort et al. 2006). For miR162 and miR168, this hypothesis might be consistent with the low levels of mature microRNA observed by

deep-sequencing, in contrast to the apparently high spliced transcript levels. For several pre/pri-miRNA loci (notably miR162 and miR168) we infer several closely related canonical introns (shared splice donors with splice acceptor sites shifted by a few tens of bases or vice-versa). We speculate that this phenomenon might be due, in part, to the incapacity of the Nonsense Mediated Decay pathway (which is dependent on ribosomal scanning of mRNAs to monitor "erroneous" splicing of non-coding transcripts (Amrani, Sachs et al. 2006).

2.6 Conclusions

We performed comparative prediction of conserved *Vitis vinifera* miRNA precursor loci, yielding over 140 high confidence predictions on the 12x genome draft. Software to assist in the design of oligonucleotide arrays for the validation of miRNA expression was developed and Oligonucleotide array and deep sequencing experiments were used to confirm the expression of mature miRNAs from most of these loci in at least one tissue or developmental stage. Many miRNAs show strong patterns of tissue specific expression. Where knowledge of the target gene for these miRNAs is available from other species, we have considered the observed expression patterns in *Vitis* to generate hypotheses regarding the physiological significance of our observations. We have shown that for many, but by no means all miRNA precursors, evidence for primary transcript expression can be obtained from high throughput transcriptome analysis, classically performed to follow expression levels of protein coding genes. Finally, we have developed a bioinformatics strategy that, when large numbers of transcriptome reads mapping to a precursor miRNA locus are available, allows the estimation of patterns of splicing and alternative splicing of pri-miRNA transcripts. Our preliminary data suggest that splicing and alternative splicing of pri-miRNAs may be a common phenomenon.

Chapter 3

Ab-initio prediction of miRNA precursors from genomic sequence data

3.1 Introduction

With the availability of a complete genomic sequence, the identification of candidate precursor sequences for members of conserved miRNA families is relatively straightforward using tools such as microHARVESTER (Dezulian, Remmert et al. 2006) or miRscan (Lim, Lau et al. 2003). However, novel or *lineage specific* miRNAs can not be identified in this way. When the current project was initiated, deep sequencing of smallRNA fractions (which has subsequently become the most common way to identify novel miRNAs) was not commonly available, and it was decided to focus on so called *ab-initio* approaches to identify *lineage specific* miRNAs in plants. Such approaches must necessarily be based on the identification of genomic sequences that, if transcribed, could fold to yield hairpins with typical characteristics of miRNA precursors as, *a-priori*, no information on the nature of the putative miRNA or miRNA* sequence is available. The situation is further complicated by the absence of well conserved primary sequence motifs associated with the specificity of DICER-like l (DCL1), the enzyme responsible for the release of the miRNA/miRNA* duplex from the hairpin precursor. Thus*, ab-initio* miRNA prediction tools must evaluate only a predicted secondary structure in order to decide if it is likely to be a valid pre-miRNA.

Simple evaluation of the energetic stability of a hairpin structure is not sufficient for the identification of plausible miRNA precursors, in a small to medium sized plant genome such as that of *Vitis vinifera* (480 megabases) a scan of the complete genome using RNALfold (Mccaskill 1990; Hofacker, Fontana et al. 1994; Ambros, Bartel et al. 2003; Hofacker 2003; Hofacker, Bernhart et al. 2004; Griffiths-Jones 2006; Meyers, Axtell et al. 2008) reveals over 4 million potentially locally stable hairpin structures with stability in the range observed for known precursor miRNAs. However, a collection of information based on primary and secondary structure characteristics might allow discrimination between potential miRNA precursors and spurious hairpins. Once a probable precursor is identified, it might be possible to identify most likely miRNA and miRNA* sequences based on their position in the hairpin and the local secondary structure of regions of the stem. Such information would be sufficient to design experimental procedures for the validation of *in-silico* predictions. Some hope for

such methods is provided by the findings that real miRNA precursors do possess some structural characteristics that distinguish them from other, non-miRNA hairpins (Bonnet, Wuyts et al. 2004; Borenstein and Ruppin 2006; Lee and Kim 2008).

It should be noted that even more than sensitivity of such a method, extremely high specificity (low false positive prediction rate) is essential to make the method useful, as even a 1% false positive rate would result in 40000 false positive predictions when all 4 million stable hairpins from the *Vitis vinifera* genome are tested for example.

Several *ab-initio* predictors of pre-miRNAs have been implemented based on context specific Hidden Markov Models (Agarwal, Vaz et al. 2010), genetic programming (Brameier, Krings et al. 2007), the identification of context robust hairpins physically close to known animal pre-miRNAs (Sewer, Paul et al. 2005), density information (Bentley, Balasubramanian et al. 2008) and Support Vector Machine (Xue, Li et al. 2005). Most of these methods rely on so-called supervised learning techniques, whereby a system is "trained" to distinguish known positives from known negative instances (in this case real miRNA precursors and hairpins that have similar overall energy, but which do not produce mature miRNAs). Due to some previous experience in the laboratory of using Support Vector Machine (Re, Pesole 2009) and encouraging results obtained in the prediction of miRNAs using this method in the literature (Xue, Li et al. 2005), it was decided to follow this approach to try to develop a reliable *ab-initio* pre-miRNA prediction tool. As with any problem in which machine learning approaches are used, the selection of characteristics or features used to describe the instances to be classified is perhaps the most important step. We have taken some features developed in previous studies and added novel descriptors of hairpin structures in order to maximize the sensitivity and specificity of our method.

3.2 General information about Support Vector Machine (SVM)

The machine learning method SVM (Support Vector Machine) is able to analyze data and recognize patterns, used for classification and regression analysis

The original SVM algorithm was invented by Vladimir Vapnik and the current standard incarnation (soft margin) was proposed by Corinna Cortes and Vladimir Vapnik (Cortes and Vapnik 1995; Vapnik 1998) (http://www.springerlink.com/content/k238jx04hm87j80g/).

Support Vector Machine performs classification by constructing an N-dimensional hyperplane that optimally separates the data into two categories

An SVM analysis finds the line (or, in general, hyperplane) that is oriented so that the margin between the support vectors (the planes that are used to orient the decision hyperplane) is maximized (Fig.3.1).

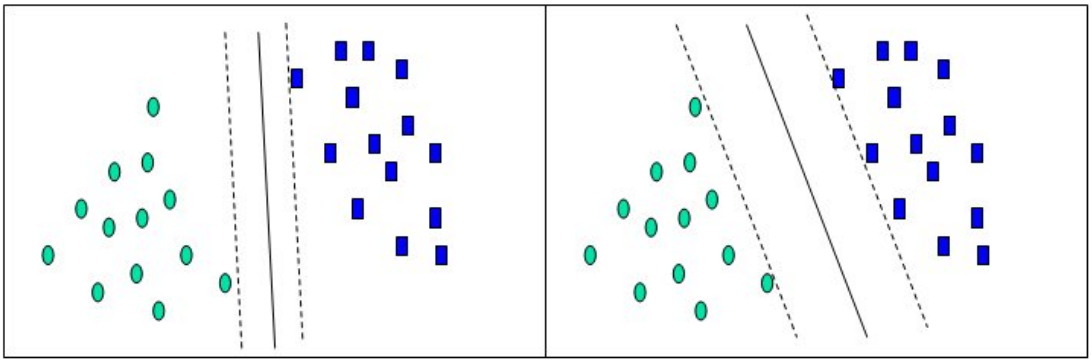


Fig.3.1 - Maximum-margin hyperplane and margins
There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the *maximum-margin hyperplane* and the linear classifier it defines is known as a *maximum margin classifier*. From http://www.dtreg.com/svm.htm

SVM has been widely applied to the prediction and classification of important biology signals such as promoters (Gordon, Chervonenkis et al. 2003), translation initiation sites (Zien, Ratsch et al. 2000) splicing sites (Zhang, Heller et al. 2003) and proteins (Leslie, Eskin et al. 2004). SVM was successfully applied to predict new virus miRNAs (Pfeffer, Sewer et al. 2005).

A classification task usually involves with training and testing data which consist of some data instances. Each instance in the training set contains one “target value” (class labels) and several “attributes” (features). The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes.

Given a training set of instance-label pairs (x_i, y_i), $i = 1, \ldots l$ where $x_i \in R^n$

$$\min_{w, b, \xi} \quad \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

subject to:

$$y_i \left(w^T \varnothing (x_i) + b \right) \geq 1 - \xi_i$$

$$\xi_i \geq 10$$

Here training vectors x_i are mapped into a higher (maybe infinite) dimensional space by the function $\varnothing$. SVM finds a linear separating hyperplane (fig) with the maximal margin in this higher dimensional space. $C > 0$ is the penalty parameter of the error term.

Furthermore, $K(x_i; x_j) \equiv \varnothing(x_i)^T \varnothing(x_j)$ is called the kernel function.

There are many kernel function, but for our analysis we consider the Radial Basis Function (RBF) kernel:

$$K(x_i; x_j) = e^{\gamma \| x - y \| 2}$$

where γ and C are parameters to train the whole training set.

We choose the RBF kernel because it nonlinearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinear.

Another reason is that RBF kernel has a minor number of hyperparameters respect to the polynomial kernel (a bigger number of hyperparameters increases the complexity of model selection).

There are some situations where the RBF kernel is not suitable. In particular,

when the number of features is very large, but it is not our case.

We have used the software LIBSVM (Chang, Lin; 2010) (http://www.csie.ntu.edu.tw/~cjlin/LIBSVM/), which provides programs for most important steps in SVM analyses and has been widely used in bioinformatics studies.

3.3 SVM workflow

An SVM analysis consists of many steps (Fig.3.2)

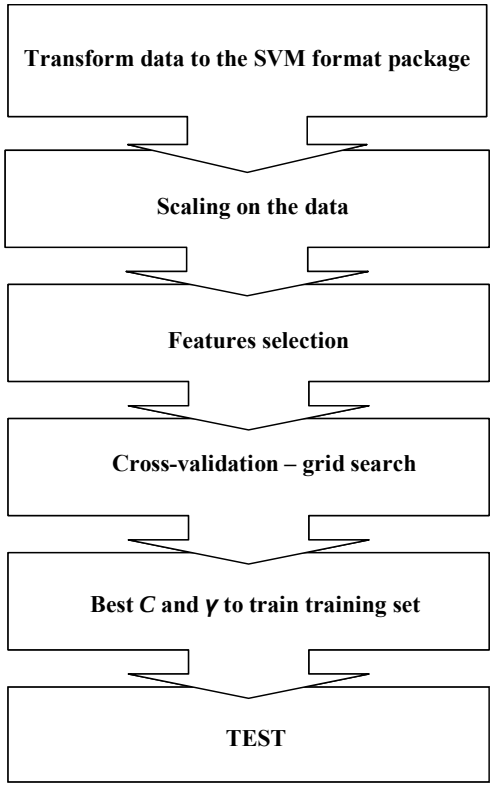


Fig. 3.2 - General workflow of a typical Support Vector Machine analysis

3.3.1 Data processing - categorical feature

SVM requires that each data instance is represented as a vector of real numbers.

Hence, if there are categorical attributes, we first have to convert the categorical attributes into numeric data. We used m numbers to represent an m-category attribute.

Only one of the m numbers is "+1", and others are "-1".

For example, a three-category attribute such as red, green, blue can be represented as [-1,-1,+1], [-1,+1,-1], and [+1,-1,-1].

3.3.2 Scaling

Scaling of data so that all attribute values fall in the same numeric range is a important step and gives several advantages.

The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges.
Another advantage is to avoid numerical difficulties during the calculation. Because kernel values usually depend on the inner products of feature vectors, large attribute values might cause numerical problems. We scaled each attribute to the range [+1; -1] relative to the largest and smallest values observed in the training set .
Thee same method is used to scale both training and testing data.
The parameter we used for scaling are l = 0 (for scaling lower 0) and u = 1 (for scaling upper 1).

We utilize scaling software "svm-scale" provided with the LIBSVM (Chang, Lin; 2010) (http://www.csie.ntu.edu.tw/~cjlin/LIBSVM/) distribution. The command line will be:

svm-scale –l 0 –u 1 –s micro_range input_features > features_scaled

where:

- –s micro_range meanings to save scaling parameters to the file micro_range (for the training set preparation). For the test set, the parameter to use is –r micro_range, that indicates to use scaling parameters from the micro_range file;
- input_features is the name of the input file that contains statistical evaluations of features set for each instance;
- l = 0 (for scaling lower 0) and u = 1 (for scaling upper 1)

3.3.3 Feature selection

In complex classification problems using multiple descriptors or features, it is usually the case that the most strong signal is derived not from the values of single features but from interactions between different features. It is also possible that several features contain more or less redundant information or that some features are positively misleading for the classification of instances. Finally, the use of excess features can, in some situations, lead to "overfitting" of the model to the training set. In this case over-estimates of the accuracy of the method can be obtained as the hyperplane generated is too specific to the behaviour of the

training set used. Thus, the selection of combinations of features that can cooperatively assist in the discrimination is a very important step.

We generated a script that automates SVM training and testing for all single features, ranking them initially for their capacity to discriminate between positive and negative instances. We then start with an SVM that uses only the most efficient feature and add the next two features in terms of how informative they are. We then exclude a single feature whose omission has minimal effect on the accuracy of the SVM. The cycle is repeated (adding 2 features and excluding 1) until all features have been incorporated into the SVM. Finally, we choose the feature combination that yields the most effective SVM. While this approach is heuristic in its selection of combinations of effective features, it is quite widely used in the literature and the best feature combination selected is usually considered to be a good indication of the best features combination possible. Only the features selected in this analysis are used in the training of the final SVM.

3.3.4 Model : Cross-validation, Grid-search and training of the SVM

There are two parameters for an RBF kernel: C and γ . It is not known beforehand which values of C and γ are best for a given problem; consequently some kind of model parameter search must be done. The goal is to identify good (C;γ) so that the classifier can accurately predict unknown data (i.e. testing data).

Overfitting occurs when a model is too specialized to the classification of a training set to be widely applied to real data. The cross-validation procedure can help to prevent the overfitting problem. In this procedure, different, randomly selected subsets of the original training set are used in an iterative training and testing cycle in order to select a model and parameters that will minimize overfitting.

Fig.3.3 represents a binary classification problem to illustrate this issue. Filled circle and triangles are the training data while hollow circles and triangles are the testing data.

The testing accuracy of the classifier in Fig.3.3a and Fig.3.3b is not good since it overfits the training data. If we think of the training and testing data in Fig.3.3a and Fig.3.3b as the training and validation sets in cross-validation, the accuracy is not good. On the other hand, the classifier in Fig.3.3d and Fig.3.3c does not overfit the training data and gives better cross-validation as well as testing accuracy.

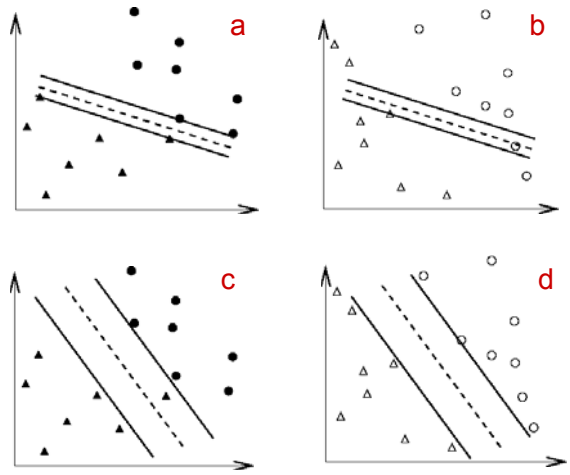


Fig**. 3.3 - An overfitting classifier (a,b) and a better classifier (c,d)**
(a) Training data and an overfitting classifier; (b) Applying an overfitting classifier on testing data; (c) Training data and a better classifier; (d) Applying a better classi_er on testing data. ▲ and ●: training data; Δ and ○: testing data).

We performed a grid-search on C and γ using cross-validation. We used the program grid.py available in the R-LIBSVM interface (http://www.csie.ntu.edu.tw/~cjlin/LIBSVM/) (Chang, Lin; 2010) which automates such a procedure. Various pairs of (C; γ) values are tried and the one with the best cross-validation accuracy is picked. The command line for this analysis is:

python grid.py features_scaled_selected > grid. features_scaled_selected

where:

- features_scaled_selected is the files scaled that contained only the selected features;
- grid. features_scaled_selected is the file that contains the C and γ values

We take the last set of values for C and γ printed and train the SVM machine.
[[local] 13 -3 98.6166 (best c=32.0, g=0.125, rate=98.9723)]

3.3.5 Output of probabilities associated with classifications

In addition to functions described until now, the LIBSVM package (Chang, Lin; 2010) (http://www.csie.ntu.edu.tw/~cjlin/LIBSVM/) allows us to train the SVM in such a way that probability scores can be associated with the classifications obtained. By including the option

"-b 1" with training and testing commands, we are able to output such scores. Thus we train the SVM with the command

svm-train –b 1 –c x –g x features_scaled_selected

where:

- -b specifies that we wish to train the machine to export probability estimates;
- -g (gamma) set the γ parameter;
- -c (cost) set the C parameter

svm-train generates a model file (features_scaled_selected.model).

3.3.6 Test phase

The test phase give us the prediction of each instances referring to model.
Is possible to evaluate a known set of instances, measuring in this way the goodness and accuracy of the machine.
Is possible to evaluate an unknown set of instances, defining the classification as real instances or false instances.
The general command line is:

svm-predict -b 1 test_set_scaled_selected features_scaled_selected.model > test_set_VS_model

where:

- -b is the probability estimates;
- test_set_scaled_selected is the set of instances scaled to classify features_scaled_selected.model is the referring model;
- test_set_VS_model is a file that contains the classification of each instance ("+1" as, real instance referring to the model and "-1" as false instance referring to the model)

The number of negative instances that are erroneously classified as positives defines the false positives (FP), while the the number of positive instances that are erroneously classified as negative defines the false negatives (FN).

In addiction, on the screen is printed the overall accuracy of the prediction, but this not separated into false positives (FP) and false positives (FN). For our purposes we are principally interested in maintaining a low false positive rate for reasons described previously. Scripts were prepared to extract this information from the output files of the LIBSVM software (Chang, Lin; 2010) (http://www.csie.ntu.edu.tw/~cjlin/LIBSVM/).

3.4 Features used to describe hairpins

We selected 72 features that we hope could represent useful information to discriminate between a precursor of miRNA and non-miRNA hairpins. Given the extreme variability of loop length in plant miRNA precursors, we use only the stem region of stem-loop structures to calculate statistics.

The first step, before calculating feature values, is extracting from the folded sequence nucleotides corresponding to the stem portion (Fig.3.4).
For this purpose we scan the secondary structure in order to find all nucleotides relationship in the structure. The scan starts simultaneously from the first symbol of the sequence (at the 5' terminus) and from the last symbol of the sequence (at the 3' terminus). So computationally we use two different index (i, j) that, at the same time, scan the sequence in direction of the loop. We checked symbols (brackets or dots) that correspond to the j position and to the i position on the secondary structure. When inverted brackets were found, two paired bases were detected; when two dots are found, a mismatch or a bugle is detected; when a dot and a bracket are found, there is a nucleotide more in one of the arms of the hairpin (we need to increase or decrease one index).
Each two nucleotides of a perfect base pairing were separated by the symbol "/" (for example G/C); each two nucleotide of an imperfect base pairing was separated by the symbol "-" (for example C-A), and in particular, in case of bugles, the absence of one nucleotide is defined with zero (for example: G-0). (Fig.4).

a

```
>ptc-MIR403c
GGUUUGUGCGUGGAUCUGAGGCCAUCACAACCGUCCACUACACGACCACCCAAUGGCUUUAGAUUCACGCACAAACUCG
(((((((((((((((((((((((((.......(((........)))......)))))))))))))))))))))))).. (-41.42)
```

b

```
>ptc-MIR403c
(((((((((((((((((((((((.......((()))......)))))))))))))))))))))))
GGUUUGUGCGUGGAUCUGAGGCCAUCACAACCGUCGACCACCCAAUGGCUUUAGAUUCACGCACAAACU
```

c

```
[G/U] [G/C] [U/A] [U/A] [U/A] [G/C] [U/A] [G/C] [C/G] [G/C] [U/A] [G/C] [G/U] [A/U] [U/A] [C/G]
[U/A] [G/U] [A/U] [G/U] [G/C] [C/G] [C/G] [A/U] [U/A] [C-A] [A-C] [C-C] [A-C] [A-A] [C-C] [C-0]
[G/C] [U/A] [C/G]
```

Fig. 3.4 – Definition of relations for each couple of bases along the stem for ptc-miR403c
(a) typical output of RNAfold : hit name, sequence and secondary structure in bracket and dot annotation with its free energy; Our script eliminates the terminal bases no paired and the loop region (b) and giveus all base relations along the stem region starting from the 5'(c). Paired bases are separated by "/", unpaired bases are separated by "-". In case of bulges, if one or more nucleotide is missed, it will be replaced by a zeo "0" (example: [C-0]).

Another step, before features calculating is to check that each hairpin presents a stable structure with a minimum free energy of folding less than -20 kcal/mol. In addition we selected a lot of hairpins that contains higher A/U contents, but allowing at least 1 G/C pairing. We know that G/C pairing is an important characteristic of all ncRNAs, but a recent discovery asserts that in miRNAs A/U pairings is higher than other RNAs and that A/U pairings show close relationship with the biogenesis of mature miRNAs, indicating that the A/U content may be a useful characteristic of miRNA (Zhang, Song et al. 2010).

3.5 Features describing sequence, secondary structure and thermodynamics of hairpins

Triplet frequencies

Given the encouraging performance of the triplet SVM classifier proposed by Xue et al. (Xue, Li et al. 2005) we started by re-implementing the features suggested in this article which consider the relative frequencies of pairing of different bases in different structural contexts in the stem region.

For any 3 adjacent nucleotides, there are 8 (2^3) possible structure compositions: "(((", "((.", "(..", "(.(", ".((", ".(.", "..(" and "...". Considering the middle nucleotide among the 3, there are

32 (4 × 8) possible structure-sequence combinations, which we denote as "U(((", "A((.", etc. This defines the triplet structure-sequence elements. Fig.3.5 illustrates how a hairpin is represented using triplet elements. We exclude the terminal loop and external single-stranded regions of the hairpin and only consider the stem portions. The number of appearance of each triplet element is counted for each hairpin (pre-miRNA or pseudo pre-miRNA) to produce the 32-dimensional feature vector. Frequencies were normalized.

In detail, we use an empty array in which each element corresponds to a specific nucleotide interaction. In Fig.3.5 there is the array we prepared counted how many times each typology of interaction occurs.

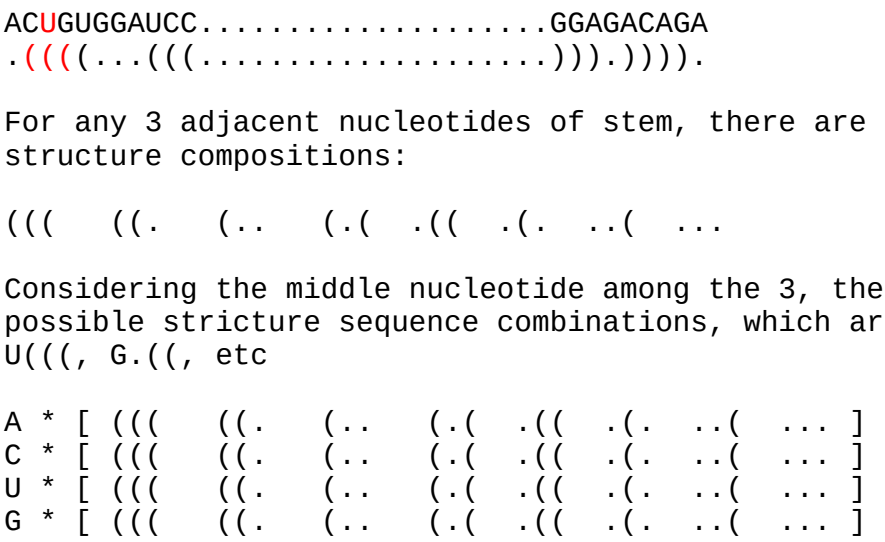


Fig.3.5 - Using the triplet elements to represent the local structure-sequence features of the hairpin. The triplet element is composed of the 3 continuous sub-structures and the nucleotide type at the middle. The appearances of all 32 possible triplet elements are counted along a hairpin segment, forming a 32-dimensional vector, which is then normalized to be the input vector for SVM Only stem portion (shadows regions) of the hairpin are computed. From (Xue, Li et al. 2005) modified

Paired and unpaired frequencies

For the stem region, we calculate the relative frequencies of all possible typologies of base interactions considering the 5' arm of the hairpin as distinct from the 3' arm.

In detail, we use an empty array in which each element corresponds to a specific nucleotide interaction (bases involved in pairings, combinations of "correctly" juxtaposed paired

unpaired bases and nature of bases involved in asymmetric bulges). We then calculate a normalized count of the frequencies of each type of interaction between bases.

The possible combinations considered are shown below, "A-C" is, for example, considered as distinct from "C-A" as the first letter always corresponds to the base found on the 5' arm of the hairpin (Fig.3.6)

```
("[A-A]","[A-C]","[A-G]",
 "[U-U]","[U-C]","[U-G]",
 "[C-C]","[C-A]","[C-U]",
 "[G-G]","[G-A]","[G-U]",
 "[A-0]","[U-0]","[C-0]",
 "[G-0]","[0-A]","[0-U]",
 "[0-C]","[0-G]","[A/U]",
 "[U/A]","[C/G]","[G/C]",
 "[U/G]","[G/U]")
```

Fig.3.6. - Array used for computing paired and unpaired bases frequencies.
Paired bases are indicated with "/", while "unpaired bases are indicated with "-".

Complementary bases frequencies

All typologies of paired bases along the stem (excluding the loop and the terminal region) were evaluated. The G/U base pairings was included, as is known to be common in miRNA precursor. This measure is distinct from those described previously as arm specificity is not considered. Frequencies were normalized to the number of paired bases.

$$bp_{A/U} = \sum_{i=1}^{n} \frac{bp_i}{n} \qquad bp_{G/C} = \sum_{j=1}^{n} \frac{bp_j}{n} \qquad bp_{U/G} = \sum_{z=1}^{n} \frac{bp_z}{n}$$

Non-complementary bases frequencies

We calculate the frequencies of 2 typologies of non complementary bases: $upb_{x\text{-}x}$ and $upb_{x\text{-}o}$ depending on the presence or not of a corresponding base on the other arm.

Thus:

$$ubp_{x\text{-}x} = \sum_{i=1}^{n} \frac{ubp_i}{n} \qquad ubp_{x\text{-}o} = \sum_{w=1}^{n} \frac{ubp_w}{n}$$

Where n is the number of unpaired bases.

Symmetry

We define one pair of non-complementary bases as a mismatch, while longer or asymmetric stretches of non-complementary bases are considered as bubbles. We developed a specific measure of bulge symmetry:

a : b

where, given n bulges, a is equal to the sum of each shorter unpaired stretches of bases (x).

$$a = \sum_{i=1}^{n} x_i$$

where, given m bulges, b is equal to the sum of each longer stretch of unpaired bases (y) .

$$b = \sum_{k=1}^{m} y_k$$

The ratio a : b belongs:

$$a : b = \frac{\sum_{i=1}^{n} x_i}{\sum_{k=1}^{m} y_k}$$

Thus for the hairpin in Fig.3.7 the calculation of ratio is a : b = 1 : 6.

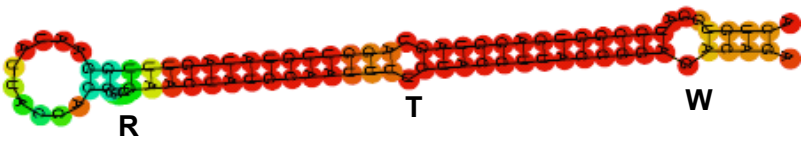


Fig.3.7 Our annotation for bulges and mismatches.
One pair of non-complementary bases as a mismatch (T), while longer or asymmetric stretches of non-complementary bases are considered as bubbles (R and W).

Thus, R and W are bulges, while T is a mismatch (according to our annotation) (Fig.3.7).
In the bulge R there are two bases in one arm of the stem and zero in the other. The bigger number of unpaired bases is 2 while the lower is 0.
In the bulge W there is 1 base in one arm of the stem and three bases in the other. The bigger number of unpaired bases is 3 while the lower is 1.
Thus the ratio value is:

$$\mathbf{a : b = (0 + 1) : (3+3)\ .}$$

The lower ratio value, more asymmetric is the stem (as consequence the precursor).
If there are no asymmetric bulges, the ratio is mathematically undefined, so we impose its value equal to 1 (maximal symmetry).

Number of paired bases for each nucleotide of the stem

Given the stem portion (excluding loop and terminals), we calculate the fraction of all bases that are involved in paired interactions

$$PB = \sum_{i=1}^{n} \frac{pb_i}{L}$$

Number of bulges for each nucleotide of the stem

We count the number of bases in the stem region not involved in paired interactions and lacking an unpaired counterpart on the other arm. The statistic isnormalized to the number of bases in the stem region.

$$\text{bulges} = \sum_{z=1}^{n} \frac{bulg_z}{L}$$

Number of mismatches for each nucleotide of the stem

Given the stem portion (excluding loop and terminals), we count unpaired bases with a corresponding unpaired base on the opposite arm of the hairpin.

$$\text{mismatches} = \sum_{j=1}^{n} \frac{mis_j}{L}$$

The value is normalized to the length of the stem

MFE (minimum free energy)

The calculation of mfe structures is based on dynamic programming algorithm originally developed by Zuker and Stiegler (Zuker and Stiegler 1981). Thus, the minimum free energy is calculated by RNAFold software by the Vienna RNA package (Mccaskill 1990; Hofacker, Fontana et al. 1994; Hofacker, Bernhart et al. 2004; Hofacker and Stadler 2006).

A secondary structure on a sequence is a list of base pairs i, j with i < j such that for any two base pairs i, j and k, l with i ≤ k holds:

$$i = k \iff j = l$$

$$k < j \Rightarrow i < k < l < j$$

The first condition implies that each nucleotide can take part in not more that one base pair, the second condition forbids knots and pseudoknots. The latter restriction is necessary for dynamic programming algorithms. A base pair k, l is interior to the base pair I, j, if i < k < l < j. It is immediately interior if there is no base pair p, q such that i < p < k <l <q < j. For each base pair i, j the corresponding loop is defined as consisting of i, j itself, the base pairs

immediately interior to i, j and all unpaired regions connecting these base pairs. The energy of the secondary structure is assumed to be the sum of the energy contributions of all loops. As a consequence of the additivity of the energy contributions, the minimum free energy can be calculated recursively by dynamic programming (Zuker and Stiegler 1981; Mccaskill 1990; Hofacker, Fontana et al. 1994; Hofacker, Bernhart et al. 2004; Hofacker and Stadler 2006).
Experimental energy parameters are available for the contribution of an individual loop as functions of its size, of the type of its delimiting basepairs and partly of the sequence of the unpaired strains. These are usually measured for T = 39 °C and 1 M sodium chloride solutions (Freier, Kierzek et al. 1986; Zhao, Yang et al. 2009; Andronescu, Pop et al. 2010).
For the base pair stacking the enthalpic and entropic contributions are known separately. Contributions from all other loop types are assumed to be purely entropic. This allows to compute the temperature dependence of the free energy contributions:

$$\Delta G_{stack} = \Delta T_{37,\,stack} - T\Delta S_{37,\,stack}$$

$$\Delta G_{loop} = T\Delta S_{37,\,loop}$$

Where ΔG_{stack} is the conformational free energy of stack and ΔG_{loop} is the conformational free energy of loop.
The structure (list of base pairs) leading to the minimum energy is usually retrieved later on by "backtracking" through the energy arrays (Hofacker, Fontana et al. 1994).
The partition function for the ensemble of all possible secondary structures can be calculated analogously.
The computation of the minimum free energy structure including the entire matrix of base pairing probabilities is considerably faster. Secondary structures are represented by a string of dots and matching parentheses, where dots symbolize unpaired bases and matching parentheses symbolize base pairs.
All MFEs were expressed as negative kcal/mol

AMFE (Adjusted MFE)

Is the MFE of 100 nucleotides and was calculated by:

$$AMFE = \frac{MFE}{L} * 100$$

Where L corresponds to the length of RNA sequence (Zhang, Pan et al. 2006).

MFEI (Minimal Folding Energy Index)

Is the MFE divided for the %G+C content (Zhang, Pan et al. 2006; Ng Kwang Loong and Mishra 2007).

$$\mathbf{MFEI} = \frac{\mathbf{MFE}}{\mathbf{C + G\ \%}}$$

AMFE rather than MFE was a better parameter to distinguish miRNAs from rRNA and mRNA. However, the AMFE of more than 50% of tRNAs falls into the range of miRNAs

RNAz- score

The significance of a predicted MFE as calculated by RNAfold is difficult to interpret in absolute terms. It depends on the length and the base composition of the sequences (longer sequences and GC rich sequences tend to have lower MFE). To some extent, the AMFE and MFEI measures described above compensate for these considerations, although a measure of the significance of an MFE score is desirable. Typically the significance of a MFE is estimated by comparing to many random sequences of the same length and base composition. If μ is the mean and and σ the standard deviation of the MFEs of many random sequences a convenient normalized measure for the significance of the native sequence with MFE is a z-score:

$$\mathbf{z\text{-}score} = \frac{\mathbf{MFE} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

The parameters μ and σ are, by construction, functions of length and base composition. However, re-sampling methods to estimate the z-score are extremely time consuming. Washietl et al. (Washietl, Hofacker et al. 2005) demonstrated that a relatively simple regression model, implemented in a SVM is capable of reliably and quickly estimating this statistic. We have taken part of the code implemented in their RNAz software for the detection of conserved RNA structures from multiple sequence alignments and customized it to allow the estimation of z-scores for MFEs of single sequences. The z score is used as a feature in our SVM for *ab-initio* miRNA discovery.

3.6 Generation of SVM

Apart from the formulation of features relevant to the discrimination of positive and negative instances and the avoidance of overfitting problems during training, the performance of machine learning methods is extremely dependent on the quality of annotation and representativeness of the data instances used in the training phase. In the case of miRNAs, this is means that it is important to have both an accurately annotate collection of real miRNA precursors to use as a positive training set and additional positive set to use in testing of the trained SVM. Additionally, we require a set of non-miRNA hairpins that would be representative of non-miRNA hairpins in the genome that we wish to study. The positive set can be obtained from databases of characterized miRNAs (for example miRBase (Griffiths-Jones 2006; Griffiths-Jones, Saini et al. 2008)) even if there is accumulating evidence that a significant proportion of sequences annotated as miRNA precursors in some species are not real miRNAs (Pam Green, personal communication). However, formulation of the negative set is more problematic as it is difficult to be confident that hairpins extracted randomly from genomic sequence are indeed non-microRNA hairpins (the annotation of microRNAs from at least most genomes is thought to be far from complete). In the development of the triplet SVM classifier, Xue et al. (Xue, Li et al. 2005) used hairpins derived from annotated coding regions, but with similar overall thermodynamic stability to known pre-miRNAs as a negative control set. The reasoning is that it is not thought that genuine miRNAs derive from coding regions in plants or animals, and so these hairpins can confidently be labelled as negative instances. Of course there is a risk that compositional or even structural constraints on coding regions might result in their characteristics not being representative of genomic hairpins in general, however, the encouraging results obtained by this approach led us to follow a similar strategy.

3.7 Initial evaluation of our machine learning strategy: Feat-SVM

Xue et al., applied Support Vector Machine (SVM) to 32 features focused on information derived from sequence and pairing of every 3 adjacent nucleotides in candidate precursors, reaching around 93% sensitivity with 12% false positives (Xue, Li et al. 2005). Our feature set includes the 32 features used in the triplet SVM classifier and many additional features described above.

3.7.1.1 *Datasets and Results*

In this section, we first describe how suitable datasets for training and testing the devised SVM strategy were prepared. Then we consider results obtained from the trained SVM.

Dataset

To obtain an initial evaluation of our approach, we trained our SVM with the same dataset used by Xue et al. (Xue, Li et al. 2005) and compared the performance with that of the Triplet SVM classifier on the same test sets as used in the initial study:

Training set

Positive: 163 real human miRNAs downloaded from the release 5.0 of the miRNA registry miRBase(Griffiths-Jones 2006; Griffiths-Jones, Saini et al. 2008). All miRNAs used in this dataset, provided by Xue et al. (Xue, Li et al. 2005) have a short loop region lacking additional predicted secondary structure (one loop structure in our terminology).

Negative: 168 hairpins derived from coding region (CDSs) of human RefSeq genes with no known alternative splice events. The CDS sequences are extracted according to the UCSC refGene annotation tables (Karolchik, Baertsch et al. 2003; Karolchik, Kuhn et al. 2008).

Test set

Positive: 69 human miRNAs downloaded by miRBase (Griffiths-Jones, Saini et al. 2008) and not included in the positive training set

Negative: 1000 hairpins derived from coding region (CDSs) of human RefSeq genes with no known alternative splice events. The hairpins was extracted from the genome region of positions 56,000,001 to 57,000,000 on human chromosome 19.

Feature selection

Feature selection was performed as presented previously. The list of selected features is provided in the first column of Tab.3.1. We note that a substantial number of the additional features included in our feature set were selected in the most effective group recovered by our heuristic procedure

Tab.3.1 Features selected by Feat-SVM and Plant-Bias SVM
With a "x" are indicated the selected features

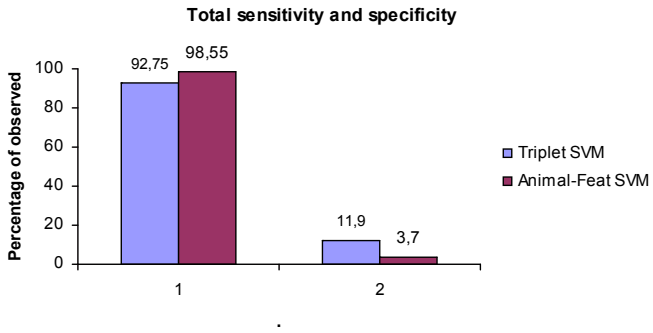
Number of feature	*Feature type*	*Feat-SVM*	*Plant-Bias SVM*
1	A...		x
2	A..(		x
3	A.(.	x	x
4	A.((	x	x
5	A(..		x
6	A(.(	x	
7	A((.	x	x
8	A(((	x	x
9	G...	x	x
10	G..(	x	x
11	G.(.	x	x
12	G.((	x	x
13	G(..	x	x
14	G(.(	x	x
15	G((.	x	x
16	G(((	x	x
17	C...	x	x
18	C..(	x	x
19	C.(.	x	x
20	C.((	x	x
21	C(..		x
22	C(.(		x
23	C((.	x	x
24	C(((	x	x
25	U...		x
26	U..(	x	x
27	U.(.	x	x
28	U.((	x	x
29	U(..		x
30	U(.(	x	x
31	U((.	x	x
32	U(((	x	x
33	A-A	x	x
34	A-C	x	x

35	A-G	x	x
36	U-U	x	x
37	U-C	x	
38	U-G	x	x
39	C-C	x	x
40	C-A	x	x
41	C-U	x	x
42	G-G	x	x
43	G-A	x	x
44	G-U	x	x
45	A-0		x
46	U-0	x	x
47	C-0		
48	G-0	x	x
49	0-A		x
50	0-U		x
51	0-C	x	x
52	0-G	x	x
53	A/U		x
54	U/A	x	
55	C/G	x	x
56	G/C	x	x
57	U/G	x	x
58	G/U	x	x
59	C-G /tot_paired_bases	x	x
60	A-U/ tot_paired_bases	x	x
61	G-U/ tot_paired_bases	x	x
62	X-0/ tot_unpaired_bases	x	x
63	X-X/ tot_unpaired_bases	x	x
64	Symmetry	x	x
65	Pairings/stem_length		x
66	Bulges/stem_length		x
67	Mismatches/stem_length	x	x
68	Stem_length	x	x
69	MFE	x	x
70	AMFE	x	x
71	MFEI	x	x
72	RNAz-score		x

Results

The graph and the table in Fig.3.8 (below) shows the accuracy of classification of test set instances by the Triplet SVM classifier and our machine (Feat-SVM).

The Triplet SVM classifier is able to classify correctly 64 of 69 real miRNA precursors (a true positive (TP) rate or sensitivity of 92.75 %).
Our SVM (Feat-SVM), correctly classifies 68 of the 69 real miRNAs (a TP rate, or sensitivity of 98.55%).



	Triplet SVM accuracy		Feat-SVM accuracy		Total instances
Positive test set	TP 92,75	FN 7,25	TP 98,55	FN 1,45	69
Negative test set	TN 88,10	FP 11,90	TN 96,30	FP 3,70	1000

Fig.3.8 – Accuracy of classification of test set instances by the Triplet SVM classifier and Feat-SVM
TP= false positives; FN=false negatives. In the x axis 1= sensitivity; 2=false positives rate

With respect to the incorrect classification of negative instances (hairpins derived from coding regions) as miRNAs, the Triplet SVM classifier incorrectly classifies 190 of 1000 hairpins as miRNAs (False Positive (FP) rate of 11.9%, or a specificity of 88.1%). Our SVM (Feat-SVM), instead classify correctly the 96,30 % of negative test set (1000 hairpins derived from CDS) so the false positive rate decreases to 3.70 %.

We conclude that our SVM improves both sensitivity and specificity of classification with respect to the Triplet SVM classifier. However, the false positive rate (3.7%) in this experiment is still too high to make the approach practical for *ab-initio* prediction of miRNAs in a large genome for reasons discussed previously (Fig.3.8)

Output of probabilities associated with classifications

In addition to functions described until now, the LIBSVM package (Chang, Lin; 2010) (http://www.csie.ntu.edu.tw/~cjlin/LIBSVM/) allows us to train the SVM in such a way that probability scores can be associated with the classifications obtained. By including the option "-b 1" with training and testing commands, we are able to output such scores. Fig.3.9 shows the distribution of probability scores for training and testing phases in the experiment described above.

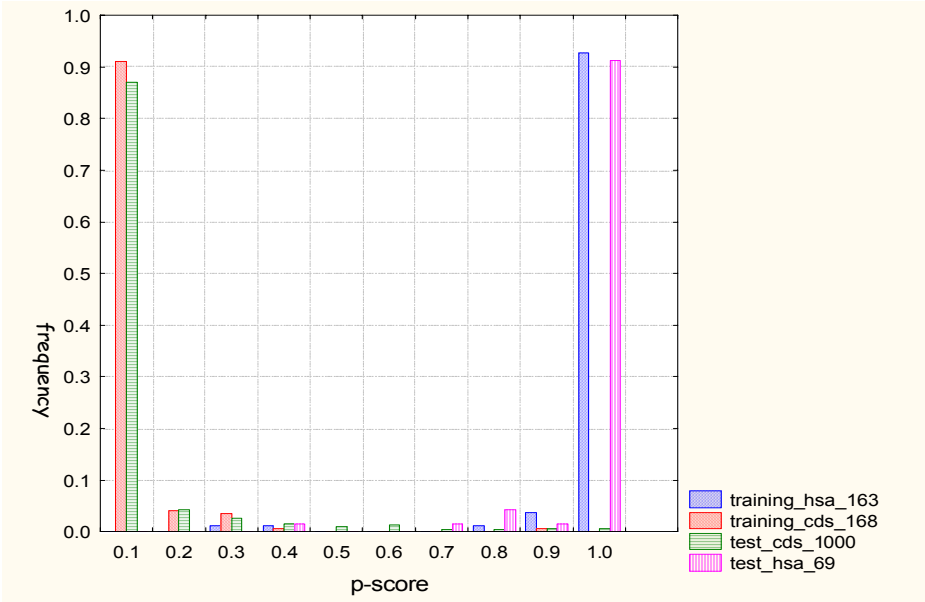
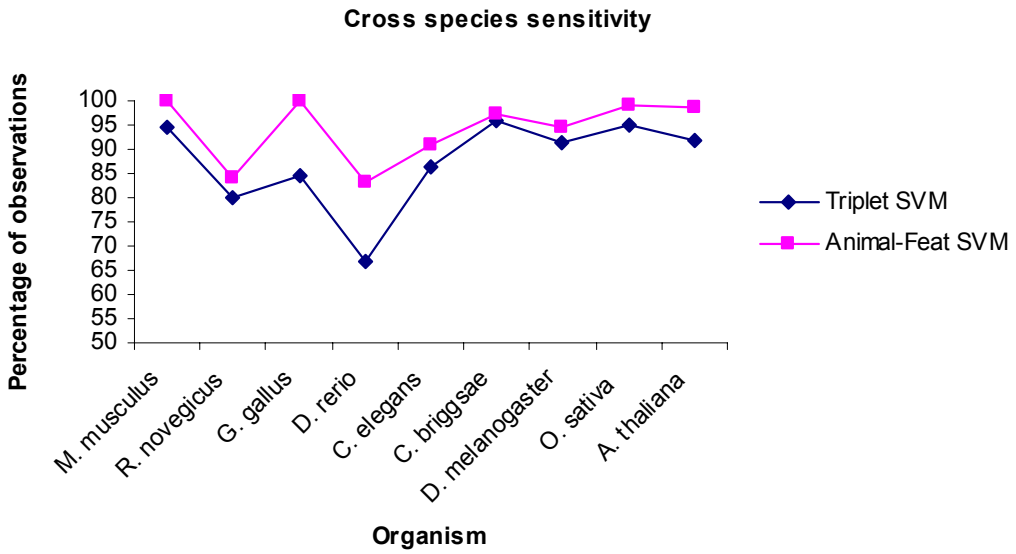


Fig.3.9 - Distribution of probability score for training and testing test of Feat-SVM
We can see that the clear majority of negative instances are associated with extremely low probabilities of being miRNAs, while the majority (>90%) of real miRNAs are classified as such with high confidence (a score greater than 0.9) (Fig.3.9). These findings give further support to the conclusion that the features we are using give good signal with respect to the status of hairpin structures, and indeed that real miRNA precursors do indeed possess distinct intrinsic properties that render them functional.

Cross species analysis

Xue at al. (Xue, Li et al. 2005) demonstrated that their Triplet SVM classifier showed similar sensitivity in the recognition of pre-miRNAs from other plant and animal genomes, even when the training had been performed exclusively on human sequences. After experimenting on the human data, we applied our SVM classifier, trained with human data, to other species to see if the additional features used retain theor value across large evolutionary distances.

The release 5.0 of the miRNA registry (Griffiths-Jones, Saini et al. 2008) contained 1138 pre-miRNAs entries from 9 species besides human (*Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Drosophila melanogaster*, *Danio rerio*, *Gallus gallus*, *Mus musculusi*, *Rattus norvegicus*, *Arabidopsis thaliana*, *Oryza sativa*).



Species	Triplet SVM	Feat-SVM	Total haipins
Mus musculus	94,4	100	36
Rattus novegicus	80,0	84,0	25
Gallus gallus	84,6	100	13
Danio rerio	66,7	83,0	6
Caenorhabditis elegans	86,4	90,90	110
Caenorhabditis briggsae	95,9	97,26	73
Drosophila melanogaster	91,5	94,36	71
Oryza sativa	94,8	98, 95	96
Arabidopsis thaliana	92,0	98,66	75

Fig.3.10 – Sensitivity of TripletSVM and Feat-SVM for the crossing species analysis

3.8 Evaluation of our second machine learning strategy: Plant-Bias SVM

For an *ab-initio* method to be suitable for analysis of large genomes, it is desirable that the sensitivity should be high, but it is essential that the false positive rate should be low, to avoid the generation of many false positive predictions when very large numbers of candidate

hairpins are tested. It has been shown that the use of imbalanced datasets (the situation when one class is much more highly represented than the other in training or testing) can effect the accuracy of SVM-based classifiers (for detailed discussion see (Batuwita, Palade 2009)). The LIBSVM software (Chang, Lin; 2010) (http://www.csie.ntu.edu.tw/~cjlin/LIBSVM/) incorporates the possibility to address this problem by specifying, a-priori, an expected relative frequency between positive and negative instances. While we do not know a meaningful value for this parameter, we performed several tests to examine the effect of a-priori class frequency weighting on the performance of our SVM, the optimal results obtained and presented below used a weighting of 10:1 in favour of non-miRNA instances. The syntax for SVM training in this case is a follows:

svm-train -b 1 -w1 1 -w-1 10 –c 32.0 –g 0.125 features_scaled_selected

where:

- -w1 is the weight of true instances (+1)
- -w-1 is the weight of false instances (-1)

Additionally, while the tests presented previously suggest that the sensitivity of our method in plants remains high, even when the SVM was trained with human data, we wished to develop a system which was trained on plant microRNA and negative training data.

3.8.1 *Datasets and Results*

In this section, we first describe how suitable datasets for training and testing the devised SVM strategy were prepared. Then we consider results obtained from the trained SVM.

Positive training set

For the positive training set, we extracted all known miRNA of *Arabidopsis* and Poplar from miRBase, the searchable database of published miRNA sequences and annotation.
For each hairpin we predict secondary structures of single stranded RNA precursor using the software RNAfold from the Vienna package (Hofacker, Fontana et al. 1994; Hofacker 2003; Hofacker, Bernhart et al. 2004).

We selected randomly 80% of all known miRNAs from *Arabidopsis* and Poplar excluding all precursors belonging to the miR160 family and obtained a total of 228 precursors, 97 from *Arabidopsis*, 131 from Poplar .

Negative training set

For the negative training set we continued to extract hairpins from coding regions as is not thought that miRNAs can originated from CDSs.

We collected all annotated coding region from the *Arabidopsis* Information Resource (TAIR), the database of genetic and molecular biology data for *Arabidopsis thaliana* (http://www.arabidopsis.org/).

For each coding region we predict secondary structures of single stranded RNA using the software RNALfold from the Vienna package (http://www.tbi.univie.ac.at/~ivo/RNA/) (Hofacker, Fontana et al. 1994; Hofacker 2003; Hofacker, Bernhart et al. 2004).

This software computes locally stable RNA secondary structure with a maximal base pair span. For a sequence of length n and a base pair span of L, the algorithm uses only O(n+L*L) memory and O(n*L*L) CPU time. We use -L 400 -d2 –noLP. Thus we "scan" all *Arabidopsis* coding regions for short RNA structures. The output consists of a list of secondary structure components of size <= L, one entry for line. Each output line contains the predicted local structure its energy in kcal/mol and the starting position of the local structure (Fig).

We prepared a script in PERL for extracting all sequences and the correspondent secondary structure from the output of RNALfold. In fact each coding region can contain several hairpin combinations.

The script, through pattern matching, finds first all “>” symbols that represent the title of the CDS region analysed. If the following lines contain symbols “(“ and “)” , the script splits each line by the whitespace character, “\s”, and inserts each element into an array (“array1”). The first element at the zero position of array is the textual representation of secondary structure, the second element at the first position of array corresponds to the minimum folding energy calculated by the software, and finally the last element of at the second position of the array is the start position of the coding region analyzed. If the following lines contains “A” or/and “C” or/and “G” or/and “U” the script splits each alphabetical letter that corresponds to the nucleotides of the entire coding sequence and then inserts each of them into an array (“array2”). Finally, the script detects the correct fraction of the coding sequence whose the

secondary structure corresponds. Scans the "array1" for finding the start position of the fraction and extracts from the "array2" the first nucleotide. Then, scan all the "array2" adding a number of nucleotides equals to the secondary structure length.

Because of the huge the number of hairpins that could derive from a coding region (3791544 in our case), we used some filters to decrease their number and select hairpins that could have very similar characteristics to real miRNA precursors. Thus, we selected all the hairpins that have a minimum folding energy smaller that -15 (Bonnet, Wuyts et al. 2004; Clote, Ferre et al. 2005) and that are longer that 70 nucleotides as suggested in published literature (Bartel 2004) reducing the number of hairpins recovered from CDSs to 49078.

The next step was to select hairpins that don't share any 20mers sequences between them. In detail, the strategy was to create two lists or array: one ("ListA") that contained all the 49078 hairpins and another, "ListB", initially empty. Thus, for each sequence of the "ListA", we checked that no 20mers of "ListB" sequences was a subsequence of hairpins of the "ListA". If so, the sequence was transferred into the "ListB" and considered for the negative training set. Because of the "ListB" was initially empty, the first hairpin sequence of the "ListA", was immediately considered for the negative training set and passed directly into the "ListB".

We choose this strategy to minimize the inclusion of real unknown miRNA families into the negative training set. We used the propriety according to which miRNAs are grouped into families and a family can be spread into different species or sometimes can be limited to a species (lineage-specific). Usually a family contains many mature miRNAs with identical or very similar sequences. Thanks to our strategy, we hope to have minimized precursors specifying miRNAs belonging to unknown miRNA families. Our final negative training set contained 2032 hairpins derived from coding regions and with features very similar to those of miRNAs.

Feature Selection

Feature selection was performed as presented previously. The list of selected features is provided in the second column of Tab.3.1. The total number of features selected by our heuristic procedure is greater than in the previous experiment. Only the frequency of U-C pairings in the stem region is excluded from both analyses.

Test set

For evaluation of the new SVM we considered several positive test sets:

- 379 real miRNAs downloaded by miRBase (Griffiths-Jones, Saini et al. 2008) and not included in the positive training set in order test also inter-plant exchangeability of the method (23 from *Arabidopsis*, 33 from *Populus*, 185 from *Oryza*, and 138 from *Vitis*);
- the miR160 family in *Arabidopsis, Oryza*, *Populus* and *Vitis* (this family was excluded from the training set, thus its detection should not be influenced by exposure to homologous sequences during training);
- 69 human miRNAs downloaded from miRBase (to perform a test of the sensitivity of the plant-trained system on animal sequences)

Following the same methodology used to generate negative training sets we generated 22456 hairpins from *Arabidopsis thaliana* CDS, 2416 from *Vitis vinifera* and 2223 hairpins from *Oryza sativa* for use in evaluation of false positive rates.

Results

The sensitivity and false positive prediction rates for the new SVM and those of the original Triplet SVM Classifier are shown in Fig.3.11-15. We note that the new SVM has a sensitivity for real miRNAs that is comparable to that of the Triplet SVM Classifier, although we detected the 98, 55 % of human miRNAs in contrast with the 92,75% of Triplet SVM and correctly classified all members of the *Arabidopsis* and poplar miR160 families that were not used in training.

However, while the triplet SVM Classifier continues to yield false positive rates in the range of 15-20% in this analysis, our plant-specific weighted classifier greatly reduces false positive predictions, both with respect to the Triplet SVM Classifier and with respect to our initial SVM. In fact, the overall FP rate of 0.52% is closer to a level that might be considered useful to perform genome-wide *ab-initio* pre-miRNA prediction. However, while a false positive rate of 0.5% seems impressive, it will still result in unacceptable levels of wrong predictions when the millions of stable hairpins predicted in a large genome are tested.

	Plants	Triplet SVM %		Plant-Bias SVM %		Total
Real plant miRNAs	*Arabidopsis thaliana*	TP 95,65	FN 4,35	TP 95,65	FN 4,35	23
	Populus trichocarpa	TP 75, 75	FN 24,25	TP 84,84	FN 15,16	33
	Oryza sativa	TP 92, 97	FN 7,03	TP 91,35	FN 8,65	185
	Vitis vinifera	TP 90,57	FN 9,43	TP 84,89	FN 15,11	138
	Tot miRNA	**TP 90,76**	**FN 9,24**	**TP 88,91**	**FN 11,09**	**379**
Hairpins from CDS	*Arabidopsis thaliana*	TN 83,42	FP 16,58	TN 99,61	FP 0,39	22456
	Oryza sativa	TN 85,65	FP 14,35	TN 98,28	FP 1,72	2223
	Vitis vinifera	TN 79,88	FP 20,12	TN 99,37	FP 0,63	2416
	Tot hairpins from CDS	**TN 83,28**	**FP 16,71**	**TN 99,48**	**FP 0.52**	**27095**

Tab.3.2. – Sensitivity and specificity of Triplet SVM and Plant-Bias SVM
TP= true positives; FN=false negatives; TN= true negatives; FP=false positives

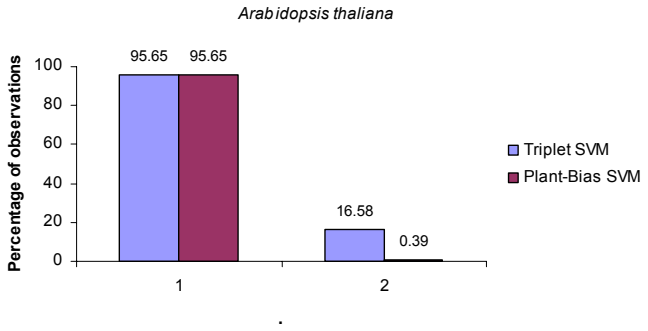


Fig.3.11 – Sensitivity and false positives rate of Triplet SVM and of Plant-Bias SVM for *Arabidopsis thaliana*
Sensitivity was calculated on 23 known miRNAs downloaded from miRBase (Griffiths-Jones, Saini et al. 2008), specificity on 22456 hairpins generated from coding regions

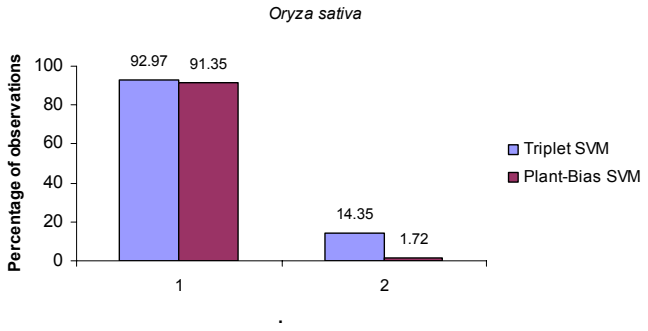


Fig.3.12 – Sensitivity and false positive rate of Triplet SVM and of Plant-Bias SVM for *Oryza sativa*
Sensitivity was calculated on 185 known miRNAs downloaded from miRBase (Griffiths-Jones, Saini et al. 2008), specificity on 2223 hairpins generated from coding regions

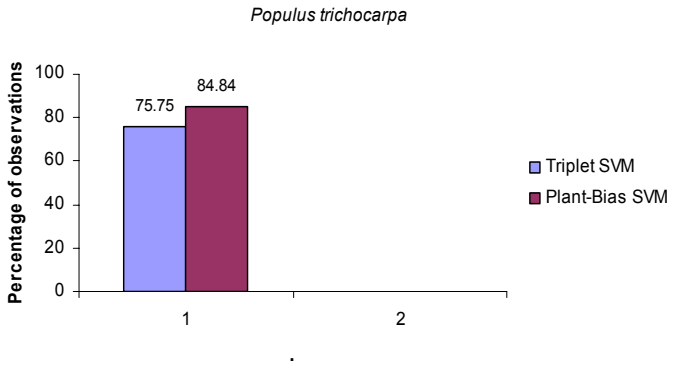


Fig.3.13 – Sensitivity of Triplet SVM and of Plant-Bias SVM for *Populus thritocarpa*
33 ptc-miRNAs was downloaded from miRBase (Griffiths-Jones, Saini et al. 2008)

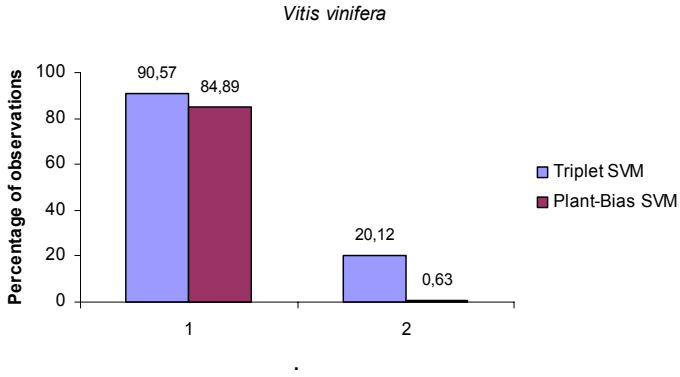


Fig.3.14 – Sensitivity and false positive rate of Triplet SVM and of Plant-Bias SVM for 138 known miRNAs and 2416 hairpins derived from coding regions of *Vitis vinifera*

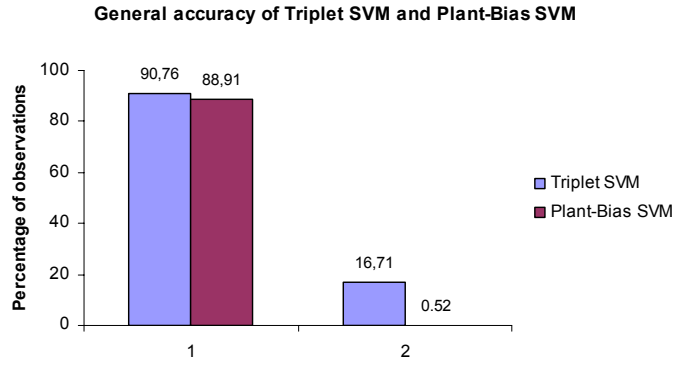


Fig.3.15 – Total sensitivity and false positive rate of Triplet SVM and of Plant-Bias SVM
Sensitivity was calculated on 379 known miRNAs downloaded from miRBase (Griffiths-Jones, Saini et al. 2008), while the false positive rate on 27095 hairpins generated from coding regions

In Fig.3.16 we can see that the clear majority of negative instances are associated with extremely low probabilities of being miRNAs, while the majority (>90%) of real miRNAs are classified as such with high confidence (a score greater than 0.9). These findings give further support to the conclusion that the features we are using give good signal with respect to the status of hairpin structures, and indeed that real miRNA precursors do indeed possess distinct intrinsic properties that render them functional.

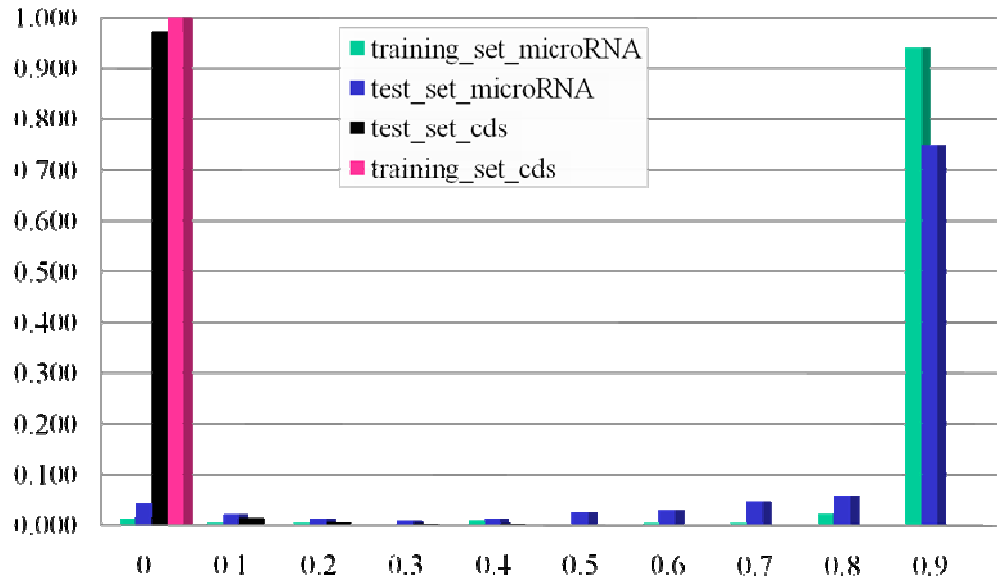


Fig.3.16. Distribution of probability score for training and testing test of Plant-Bias SVM

3.9 Conclusions and future directions

We have demonstrated that the performance of existing SVM-based pre-miRNA classifiers can be substantially improved by the addition of extra features, through careful feature selection and with the use of class frequency weighting approaches. The minimization of false positive rates necessary for the implementation of whole genome *ab-initio* miRNA prediction, comes, in our hands, at the expense of a small loss of sensitivity. It should be stressed that despite our best efforts, we can not be completely sure that all instances used in training or testing (positive or negative cases) are indeed correctly annotated. This is to say that the results obtained could be a slight under-estimation of the performance of the classifier if some instances labelled as real miRNAs are not really functional, or vice-versa.

Despite the improvements in specificity demonstrated here, we do not yet consider the method presented here to be suitable for whole genome scans. With 4 million input hairpins (the number of hairpins of length at least 70 bases with MFE < -15 kcal/mol that we detect in a genome-wide scan of *Vitis vinifera*), we might still expect to recover around 20,000 false positive predictions. This number is too high to be used to direct manual or even moderately high throughput validation experiments.

In the last 2 or 3 years, efforts towards detection of lineage-specific miRNAs have, in general, shifted away from *ab-initio* prediction strategies and towards the study of ultra-high throughput smallRNA sequence data (see introduction and Chapter 4 of this thesis). In this latter type of approach, smallRNAs are mapped to a reference genome, and pairs of sequences that could represent miRNA/miRNA* duplexes are used to define novel miRNAs. Despite it's non-applicability to perform genome-wide *ab-initio* miRNA predictions, we have incorporated our classifier into an analytical pathway for the detection of miRNAs from deep sequence data, and as shown in the next chapter, probability scores from the SVM can be used to rank predictions from deep sequence data for experimental validation.

Chapter 4

MiRNA/miRNA* detection and prediction of putative *lineage specific* miRNAs

4.1 Introduction

Ultra deep sequencing of plant smallRNAs offers the possibility to identify novel and potentially *lineage specific* miRNAs. However, given the huge volume of data generated, and the fact that many or even most smallRNA sequences do not correspond to miRNAs, but to other small interfering RNAs or even to degradation products of other RNAs, it is necessary to use specific bioinformatics methods in the interpretation of such data. Clusters of reads that map on the genome with the patterns expected for processed precursors (strand specific, predominant miRNA and miRNA* species) can indicate the presence of miRNA genes without obvious homology to annotated miRNAs. However, exhaustive analysis of all possible secondary structures involving all reads mapping close together on the genome would be extremely computationally demanding. Several bioinformatics approaches have been used to efficiently filter large quantities of NGS smallRNA data, and to discover loci showing typical patterns of miRNAs (Friedlander, Chen et al. 2008; Moxon, Jing et al. 2008).

In miRDeep (Friedlander, Chen et al. 2008) and miRCat (Moxon, Schwach et al. 2008), which are widely used for this purpose, reads are mapped to the reference genome, and typically the reads mapping to annotated structural RNAs (tRNA, rRNA snoRNA etc) and highly repetitive regions are excluded. The next filter is to identify regions with 2 very short and well defined peaks of read density close together on the same strand with a low number of reads deriving from the antisense strand (potential miRNA/miRNA* pairs). Only if these strict definitions are met will the relatively slow step of sequence extraction and secondary structure prediction be performed. This last step is important to ensure that the candidate miRNA/miRNA* pair can indeed form a duplex in a predicted precursor structure that conforms to some minimal energetic parameters. Such a strategy can be optimized to quickly search for loci that conform to relatively stringent criteria for miRNA gene annotation (Meyers, Axtell et al. 2008).

However, it is thought that younger (often novel and *lineage specific*) miRNA genes may typically by processed in less precise ways than the highly conserved miRNAs that are more

characterized in the literature and used to establish the "expected" patterns of miRNA processing. Such miRNA genes might show atypical hairpin structures, imprecise cleavage, or unusual miRNA/miRNA* lengths. Furthermore, several miRNA precursors are known to produce more than one mature miRNA/miRNA* pair (Bologna, Mateos et al. 2009; Schwab and Voinnet 2009). This type of precursor cleavage pattern is likely to be difficult to detect with miRDeep as peaks of read mapping density corresponding to the miRNA and the miRNA* will be too long to be recognized as miRNA-like patterns.

4.2 An alternative approach for the detection of novel miRNA precursors with high throughput smallRNA sequence data.

We wished to develop a method for the detection of novel miRNA precursors that would be able to detect precursors that are processed to give canonical patterns of miRNA/miRNA* products (Kurihara and Watanabe 2004; Meyers, Axtell et al. 2008), but also precursors that produce multiple stable products and precursors that produce significant quantities of non-specific smallRNA products (Bologna, Mateos et al. 2009; Schwab and Voinnet 2009). We thus wished to avoid using the detection of discrete peaks of tag mapping density on the reference genome as the first filtering step. As an alternative, we developed a rapid *in-silico* assay to determine whether any two tags could possibly constitute a miRNA/miRNA* pair without extracting genomic sequence information and modelling potential secondary structures. The trick we used was to search for pairs of reads that could form complementary duplexes with a stretch of at least 7 complementary bases between the 2 reads generating a duplex with potentially 3' overhangs of one or two bases. This is extremely fast to perform and means we can exclude most irrelevant candidates before modeling the secondary structure of potential precursors. We do not attempt to establish the most stable complimentary structure between two reads, but we simply juxtapose the sequence of the read that maps in the upstream position of the genomic sequence with the reversed sequence of the second read in ways that would allow 3, 2, 1 or 0 base 3' overhangs if the sequences were complimentary in a hairpin structure (Fig.4.1). In this way, we do not need to worry about potential short asymmetric bulges in the miRNA/miRNA* structure but we can quickly identify pairs potentially capable of forming duplexes with extensive complementarity. Observation of the patterns of miRNA/miRNA* interactions in characterized miRNAs from miRBase and empirical testing of different parameter settings led us to accept only pairs with at least 7

consecutive paired bases. These settings maximize sensitivity of the analysis with respect to known miRNAs while minimizing the amount of false positive pairs that must be modeled for secondary structure and other characteristics.

In addition to this test, we employed several other more standard filters to eliminate candidate loci that were unlikely to represent real miRNA precursors. The pipeline that was implemented is described schematically in Fig.4.1.

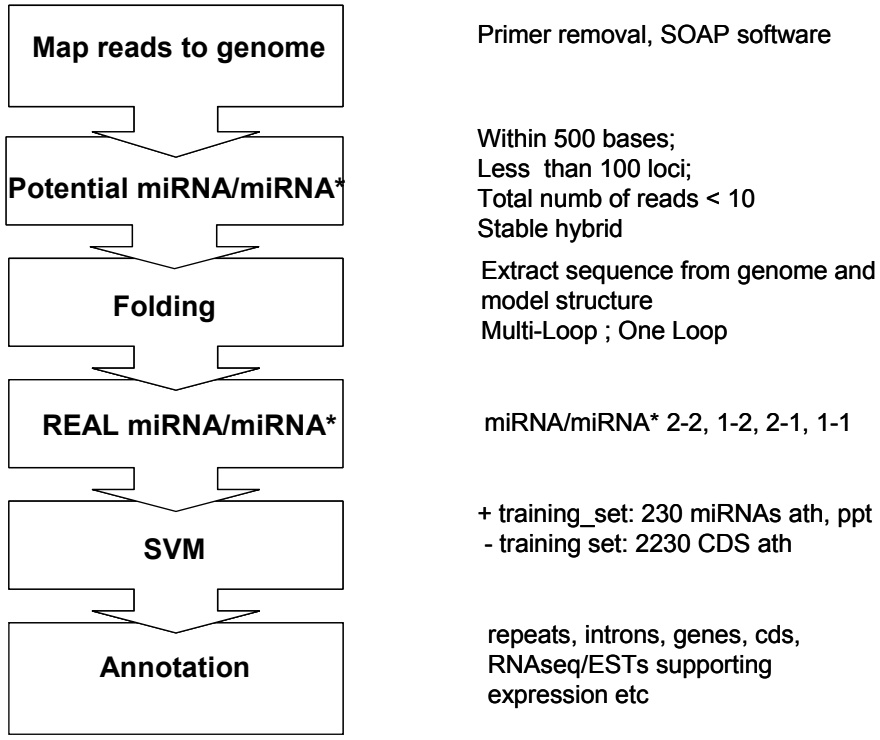


Fig. 4.1 – Pipeline for teh prediction of new miRNAs.
The pipeline is divided into several steps. After mapping tags into the genome (Map reads to genome), we selected all putative miRNA/miRNA*that can yield hybrids with at least 7 consecutive paired bases to define genomic coordinates of putative miRNA precursor sequences (Potential miRNA/miRNA*. After folding each structure (Folfing), was selected all miRNA/miRNA* that presents different types of overhangs at their 3' ends (2-2 = perfect miRNA/miRNA*, while 1-2, 2-1 , 1-1 = imperfect miRNA/miRNA*).Each hairpin was evaluated by a classificatory (SVM=Support Vector Machine). With the SVM score and the integration of other information (annotation into genome, expression level..) selected miRNA candidates are ranked.

In summary: adapter sequences added to smallRNAs during preparation for sequencing were removed as described in previous chapters using custom scripts. Identical reads were merged into clusters and the frequency with which each sequence was observed was recorded. Unique sequences were mapped tags into the genome with the program SOAP (Li, Li et al. 2008) accepting only perfect matches. The first criteria was that the distance between each pair of

tags mapped on the genome could not be bigger that 500 bp, as this corresponded to the maximal length of a precursor of a well characterized plant miRNA from miRBase (Griffiths-Jones, Saini et al. 2008) when the analysis was performed. In addition, we take into account only tags that do not map perfectly to more than 100 genomic loci in order to exclude reads deriving from extremely repetitive regions and which are unlikely to represent true miRNAs. Finally, we consider as candidate miRNA/miRNA* pairs only pairs of sequences where the total frequency of reads observed for the pair was greater than 10. This filter excludes comparisons between pairs of reads that are observed only extremely infrequently and would be unlikely to correspond to peaks of density detected by other methods. Pairs of reads that follow the above rules are subjected to the rapid test for potential miRNA/miRNA* pairs that was previously described.

Pairs of reads that can yield hybrids with at least 7 consecutive paired bases are used to define genomic coordinates of putative miRNA precursor sequences (Fig.4.2).

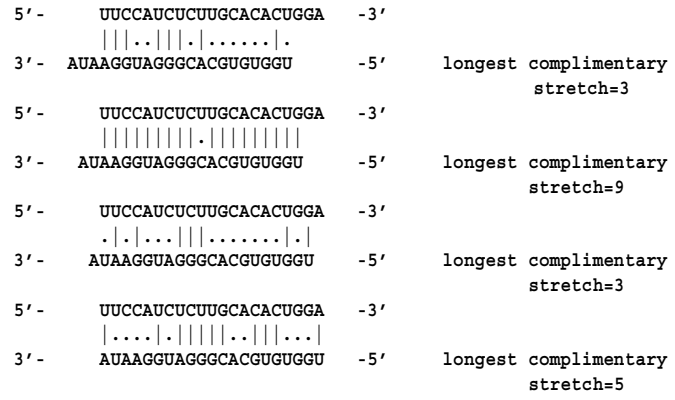


Fig. 4.2 – Generation of the stable hybrid.
Example of a potential miRNA/miRNA* duplex. In this duplex, the pairing that allows the selection is the second in which at least 7 nt are paired.

For this step, we rely on the observed tendency for miRNA/miRNA* pairs to be often situated towards the base of pre-miRNA hairpin structures (Bernstein, Caudy et al. 2001; Lee, Kim et al. 2004; Vermeulen, Behlen et al. 2005), as with comparative pre-miRNA prediction tools such as microHARVESTER (Dezulian, Remmert et al. 2006) once the genomic coordinates of a candidate miRNA/miRNA* pair are known, extensions of 10 bases 5' and 3' of the limits of the miRNA/miRNA* pair are used to define putative precursor sequences. Even errors in this assumption are unlikely to be problematic as it is known that miRNA precursor structures

tend to assume similar secondary structures even when the precise coordinates of the hairpin are changed, unlike non-miRNA hairpin structures (Bonnet, Wuyts et al. 2004; Borenstein and Ruppin 2006; Lee and Kim 2008)

The genomic sequence of a putative pre-miRNA structure was extracted and the global most stable secondary structure was estimated using RNAFold (Mccaskill 1990; Hofacker, Fontana et al. 1994; Hofacker and Stadler 2006). RNAfold exports a textual representation of the predicted secondary structure where each bracket corresponds to a base-pairing, in particular "(" pairs with ")", while each dot "." corresponds to a base not paired (it could be mismatch or a base in excess). For example loops are represented uniquely by dots, while stems contain a variety of base parings, bubbles, mismatches (Fig.4.3).

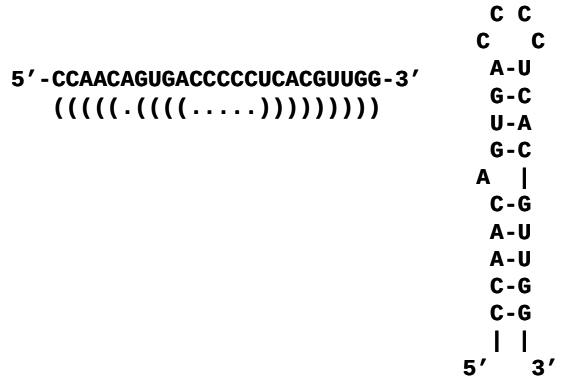


Fig. 4.3 - Textual representation of secondary structure, according to RNAfold.
Each bracket corresponds to a base-pairing, in particular "(" pairs with ")", while each dot "." corresponds to a base not paired (it could be mismatch or a base in excess).

In order to allow rapid evaluation of predicted secondary structures, we developed a script in PERL able to scan all base pairing possibilities using the textual representation of secondary structure. The textual secondary structure could be very complex, as in case of loop structures that contain additional secondary structure elements (we use the term multi-loop hairpins). Because a mature miRNA (with its corresponding miRNA*) is usually located at the beginning of the hairpin (or base of the stem) (Kim 2005; Zeng, Yi et al. 2005; Seitz and Zamore 2006), we transform each multi loop hairpin into a single loop one by conceptually dissolving secondary structures that fall within the "loop" of a potential stem structure defined by the positions of the putative miRNA and miRNA* sequences (Fig.4.4). For the analyses we perform, in practice, we don't need structural information for other part of the hairpin with

the exception of the stem nearest to the 5’ and 3’ terminus of the predicted structure. A multi-loop hairpin thus becomes a single stem with a long loop (single-loop hairpin).

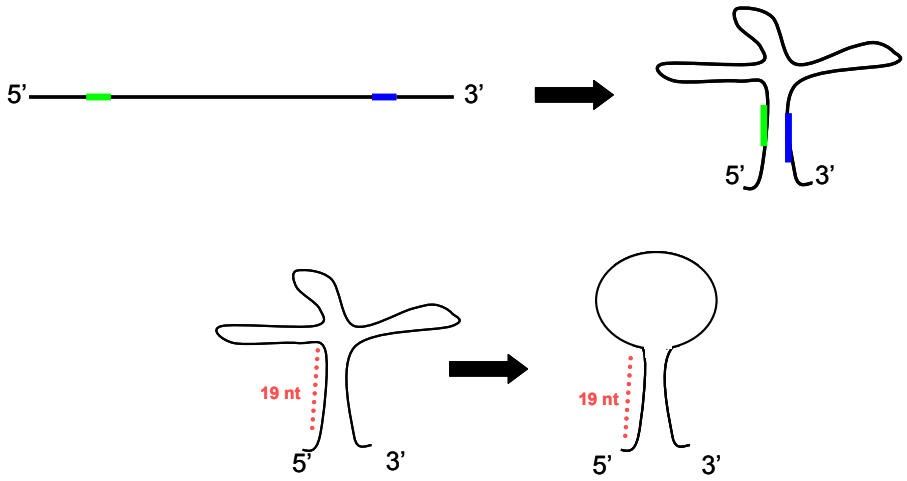


Fig. 4.4 Example of how we manage a multi-loop hairpin.
miRNA/miRNA* is located at the base of the stem (Kim 2005; Seitz and Zamore 2006) and usually the stem is bigger than 19 nt. Taking into account this criteria, we transform each multiple hairpin into a single loop hairpin.

The script we developed is articulated and complex. It scans the hairpins starting simultaneously from the 5’ and from the 3’ termini, finding all correct base pairings (defined as “(“ and “)”), and all mismatches (represented as positionally corresponding pairs of “.”) or all bubbles (defined as one or more dot “.” without corresponding bases in the opposite arm). When an inversion of brackets is found, another loop is begins. To distinguish all hairpins’ ramifications, each stem is defined by a number. In this way, the first stem corresponds to number zero, the second to the number one and so on (Fig.4.5).

```
>ath-MIR166b MI0000202 Arabidopsis thaliana miR166b stem-loop
UUUGGAGGGGGGGACUUGGUUGGUUCCUUGGGCUUCCGGAGGGGAUCCUUCCUUUAAUUUUCCUUGAAGAUUACCAGAUGUCCUUGCAAGGAUCCUUGGAUUGGAUUUUGGAUUUUAAGGGGGGUUUUUAAGUGGGUUCCGUUCCC
.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(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)0).0)0)0)0)0)0)0)0)0)0)0)0)0)0)...0)0).0)0)0).0)0)0)0)

>ath-MIR166c MI0000203 Arabidopsis thaliana miR166c stem-loop
2G2C2G2A2UU2U2A2GU2G2U2U2GA2GA2G2G2AU2U2GUU2G2U2C2U2G2GCU2C2G2A2G2G2U2C2A2UGA2A2GAAGGGAAGAAUCAOCGUUCGAA1U1U1A1A1U1U1U1GGAAGAA1C1A1A1A1U1U1A1AGAAAACC2C2UAG2A2U2G2AU2U2C2U2C2C
2(2(2(2(2(.2(2(2(.2(2(2(2(.2(2(.2(2(2(.2(2(..2(2(2(2(2(2(2(..2(2(2(2(2(2(2(2(..2(2(..0(0(0(.....0)0)0)...1(1(1(1(1(1(1(1(1(1(.....1)1)1)1)1)1)1)1).....2)2)..2)2)2)2).2)2)2)2)2)

>ath-MIR166d MI0000204 Arabidopsis thaliana miR166d stem-loop
2G2G2U2U2GA2GA2G2G2A2A2UAUU2G2U2C2U2G2GCU2C2G2A2G2G2U2C2A2UGA2A2GAAGGOAOUOCGGUAOGOAOUUOUGA1U1U1C1A1U1U1UUAAA1G1A1G1U1G1A1AAUCC2C2UAA2A2U2G2AU2U2C2U2C2GGA2C2C2A2G2G2CUUC2A2U2U
2(2(2(2(2(.2(.2(2(2(2(2(2(...2(2(2(2(2(2(2(..2(2(2(2(2(2(2(2(2(2(..2(2(..0(0(0(0(....0)0)0)0)..1(1(1(1(1(1(1(....1)1)1)1)1)1)1)....2)2)..2)2)2)2)2).2)2)2)2)2)2)2)..2)2)2)2)2)2)2)...2)2)2)

>ath-MIR166e MI0000205 Arabidopsis thaliana miR166e stem-loop
U2U2G2A2G2G2G2G2A2A2U2GUU2G2U2C2U2G2GCA2C2G2AG2GC2CC2U2U2A2A2C2UUOAOGOAOUCOUOAOUOAOUUUGAUUAUAUOAOUOAOUOGOUOCOUCUUCU1U1U1A1U1U1C1A1U1U1AGUCUA1U1AC1A1U1G1A1A1U1G1AUCAUUUUAC
.2(2(2(2(2(2(2(2(2(2(..2(2(2(2(2(2(2(..2(2(2(.2(.2(.2(.2(2(2(2(2(2(.0(0(0(0(.0(0(0(0(0(..........0)0)0)0)0).0)0)0)0)0).....1(1(1(1(1(1(1(1(1(1(.....1)1).1)1)1)1)1)1)1).....
```

Fig. 4.5 – Our new nomenclature for secondary structure
In order to distinguish all structural ramifications, each stem is defined by a number. In particular, each nucleotide (or symbol) is preceded by a number that refers to a specific stem. Thus, the first stem corresponds to number zero, the second to the number one and so on. One-loop hairpin will be characterized only by the stem number zero. Only a part of sequence is shown.

Simple filters were employed to exclude short stems (that are not plausible to appertain to a miRNA precursor). We require that the outermost stem must be greater than 19 nucleotides and if not, we consider the next stem. If it respects this length criterion we consider it as the outermost stem. In the presence of many complex structures we must scan the hairpin multiple times until we finished pairing all nucleotides.

The script generates a graphical representation of the hairpin in HTML format (Fig.4.6). Thus, each stem is displayed in a different colour in order to easily distinguish the ramification levels of the hairpin. In addition, a txt file with a new textual representation of secondary structure is generated. Now we are able to detect the exact structural juxtaposition of pairs of tags within the structure and determine whether the candidate sequences can form a plausible miRNA/miRNA* pairing.

Fig. 4.6 – Graphical representation of each hairpin in HTML format
Each stem is represented with different colours in order to easily distinguish the ramification levels of the hairpin. Only a part of sequence is shown.

For each textual representation of secondary structure all stem base pairing was positioned into an array. In this way, the first nucleotide of the sequence and its correspondent base pair are, both, at the first position of the array, the second and the second-last one at the second position and so on, while unpaired bases and asymmetric bulges are depicted in an analogous way. Any presence of bulges was evaluated. In detail, in the array we use an ad-hoc representation of nucleotide interactions. Each two nucleotides of a perfect base pairing were separated by the symbol "/" (example G/C); each two mismatched nucleotides were separated by the symbol "-" (example C-A), and in particular, in case of bugles, the absence of one

nucleotide is defined with zero (example: G-0). Given the relation "C-A", the nucleotide "C" belongs to the tag-sequence nearest to the 5' terminus, while the nucleotide "A" belongs to the tag-sequence nearest to the 3' terminus.
With this information, we are able to select all potential miRNA/miRNA* pairs of tags that form a canonical miRNA/miRNA* duplex.

If we know the mature miRNA sequence and want to detect its correspondent miRNA*, first we should map the sequence of mature into the secondary structure and then shift two nucleotides in order to find the start position of the miRNA* according to the known patterns of dsRNA cutting performed by enzymes of the DICER family (Fig.4.7) (Bernstein, Caudy et al. 2001; Lee, Ahn et al. 2003; Lee, Nakahara et al. 2004). Thus, scan the secondary structure adding a number of nucleotides equal to the length of the mature sequence. We developed a script that uses the same trick, but, as we did not know which tag of the pair corresponds to the sequence of mature or to the sequence of the star (miRNA*), we considered the tag pair nearest to the 5' terminus as *reference tag* , then mapped into the genome and found the its correspondent sequence at the 3' terminus shifted of two nucleotide. We checked that this sequence is identical to the remaining sequence of the pair: the *complementary sequence*. If the *complementary sequence* is the same, we consider the pair of tags as an exact miRNA/miRNA* duplex.

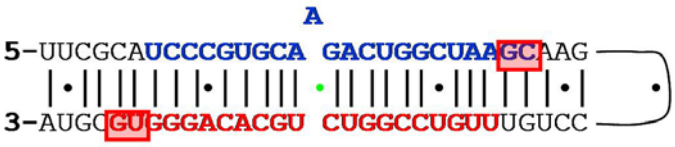


Fig.4.7 - Example of *overhangs* on the 3' end in a pre-miRNA sequence
We use the characteristic of 2-nucleotide *3'* overhang in order to detect the miRNA* sequence of each mature miRNA

In detail, given the base pairs relation "C-A", the nucleotide "C" refers to the *reference sequence*, while the nucleotide "A" refers to its *complementary sequence*. Given the relation "0-A", there is not a corresponding base in the 5' terminus (there is a bulge in the 3' arm terminus). In this case the script will take into account a nucleotide more for the definition of the *complementary sequence* during the scan.

In case of a bulge at the first position of the *complementary sequence*, we look for the first nucleotide paired in a previous position of the array and consider this as the start position of the *complementary sequence.*
In case of bulges at the last position of the *complementary sequence*, we look for the next paired nucleotide of the array and consider it as the stop position of the *complementary sequence.*

Each nucleotide of the *complementary sequence* was inserted into an array until the end of the secondary structure scan. After reversing the order of the elements (nucleotides) of the array, each nucleotides, from the first until the last position of the array, were recovered in order to obtain the *complementary sequence*.

In order to identify the complementary sequence, it is necessary to consider carefully the structure of the candidate miRNA/miRNA* hybrid as more simple strategies assuming that the complementary sequence will have the same length as the reference sequence can be misled in situations where asymmetric bulges are present in the hairpin as demonstrated in Fig.4.7 where a 1 nucleotide bulge, leads to an error in the estimation of the complementary sequence using a simple approach.

Enzymes of the DICER family tend to cut dsRNA to yield duplexes with 2base 3' overhangs (Bernstein, Caudy et al. 2001; Lee, Ahn et al. 2003; Lee, Nakahara et al. 2004; Vermeulen, Behlen et al. 2005). However, as mentioned previously, we wished to detect precursors that may not exactly follow canonical patterns of mature miRNA biogenesis. In addition, we reasoned that if miRNA/miRNA* excision was not precise, for miRNA/miRNA* pairs expressed at low levels, we might not always detect exactly canonical miRNA/miRNA* sequence pairs. Accordingly, we decided to include in our analysis some imperfect pairs of tags. In particular we considered cases in which the 3' termini of the miRNA/miRNA* hybrid did not present canonical 2 base overhangs. We developed a simple nomenclature to describe these situations: 2.1 (missed one nucleotide in the 3' overhang of the 3' read), 1.2 (misses one nucleotide in the 3' overhang of the 5' read), 1.1 (single base overhangs at the 3' termini of both reads considered) (Fig.4.8).

Despite these relaxed definitions of candidate miRNA/miRNA* juxtapositions, we also require that the miRNA/miRNA* hybrid follows some simple characteristics typical of

known pre-miRNA structures. Firstly that not more than 4 bases in either of miRNA/miRNA* sequence candidates are not paired in the secondary structure. Secondly that no more than 2 consecutive bases should be unpaired and thirdly that asymmetric bulges should not be more than 2 bases in length. Finally, we require that the minimum free energy of folding for a valid secondary structure should be less than -25 kcal per mole. These parameters are derived from empirical studies of pre-miRNA structures in plants and are similar to filters implemented in comparative miRNA prediction software such as microHARVESTER (Dezulian, Remmert et al. 2006) or MirCheck (Lai, Tomancak et al. 2003).

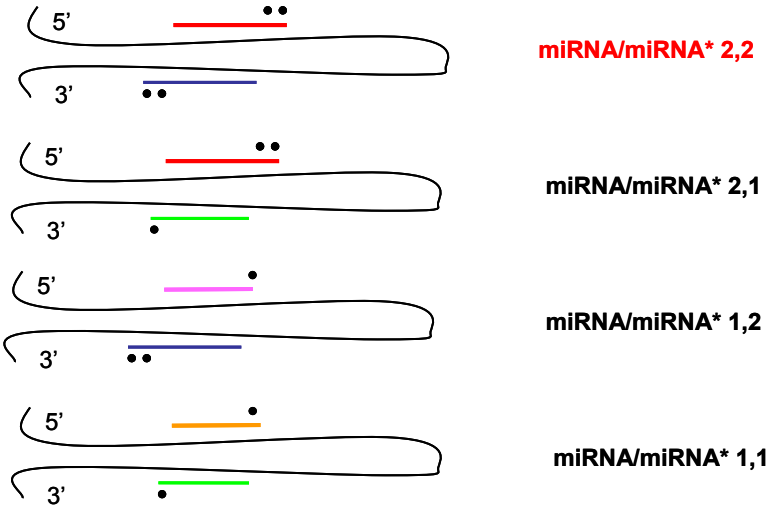


Fig. 4.8 – General scheme for miRNA/miRNA* selection.
Dots indicate the number of overhanged nucleotides taken into account for the selection (2 dots = 2 nt). We use a simple nomenclature to describe these situations: 2.1 (missed one nucleotide in the 3' overhang of the 3' read), 1.2 (misses one nucleotide in the 3' overhang of the 5' read), 1.1 (single base overhangs at the 3' termini of both reads considered).
The condition 2.2 indicates a miRNA/miRNA* duplex that respect the classical criteria described in literature (Bernstein, Caudy et al. 2001; Lee, Ahn et al. 2003; Lee, Nakahara et al. 2004; Vermeulen, Behlen et al. 2005).

After identification of all genomic loci that produce pairs of reads fulfilling all the aforementioned criteria and which could represent possible DICER products in terms of their orientation on modelled secondary structures, we perform a simple positional clustering to merge overlapping and nested hairpins, where the largest locus for such clusters of positions inherit information regarding potential miRNA/miRNA* pairs for nested and overlapping loci. This step is important for several reasons, firstly to reduce the number of candidate loci, secondly because many miRNA precursors can produce more than one distinct

miRNA/miRNA* duplex and because our strategy of allowing candidate hybrids with imperfect 3' overhangs means that where reads shifted by one or two bases with respect to each other are present, both correct and slightly incorrect miRNA/miRNA* pairs can be recovered for the same locus.

Once redundant loci that produce more than one plausible miRNA/miRNA* sequence have been merged, several additional statistics and information sources are consulted to provide additional evidence regarding the possibility that the locus could represent a real miRNA precursor. First of all, we recover all smallRNA reads that map to the defined locus, either on the sense or antisense strand with respect to the prediction. A minimum of antisense smallRNAs and the presence of well defined peaks of read density around the predicted miRNA/miRNA* sequences is consistent with miRNA biogenesis rather than other classes of siRNA. Secondly, the genomic coordinates of the locus are compared to annotations of coding genes,, structural RNAs and repetitive elements/transposons. Real miRNAs do not tend to derive from coding regions, from structural RNAs or from transposable elements. Third, we attached information regarding evidence for primary transcription of the locus from RNAseq experiments, evidence for transcription and particularly tissue specific transcription can provide support to a prediction.

We also use the Support Vector Machine (SVM) that was previously developed as part of an *ab-initio* miRNA detection strategy (see Chapter 4) to evaluate candidate pre-miRNAs discovered through the deep sequencing analysis pipeline described here. Scores from the SVM analysis are attached, along with expression and annotation data to each precursor prediction.

We developed a script able to integrate all this information in a single file. In Fig.4.9 the standard output of the program is shown.

In practice, we use the SVM score to rank predictions and exclude predictions from coding regions, annotated transposons and structural RNAs. In accordance with popular criteria (Meyers, Axtell et al. 2008) and expectations about the biogenesis of miRNAs not involving antisense transcription or RNA-dependent RNA polymerases, we also excluded predictions where more than 10% of the total smallRNA reads derived from the opposite strand Empirical studies suggest that, at least with the data considered here, this strategy minimizes loss of good predictions and maximizes exclusion of spurious predictions. However, all

parameters used in our pipeline can be modified by the user to impose more strict or more relaxed analyses and final results can be ranked according to a variety of criteria.

Of course, high throughput analytical pipelines such as the one described here can be extremely effective in the identification of miRNA precursor loci from NGS smallRNA deep sequence data, however, we consider that manual examination of candidates, coupled with eventual target prediction and validation are essential steps for conclusive demonstration of biological significance.

```
➢vvi-MIR160e
UAAGCAUAUAUGCCUGGCUCCCUGUAUGCCAUUUGCAGAGCCCACCGGCACAUCGAUGGCCUUCGUGGAUGGCGUAUGAGGAGCCAUGCAUAUGCCCCAUCUG
...((((((..(((.(((((((..((((((((((..(..((.(((.(((....))).))).))..)..)))))))))..)))))))).)))))))).......
(-51.10)
          UGCCUGGCUCCCUGUAUGCCA    254_143 3                    GCGUAUGAGGAGCCAUGCAUA  57_9 1
          UGCCUGGCUCCCUGUAUGCC      6_2 5                    UGGCGUAUGAGGAGCCAUGCA     1_0 1
          UGCCUGGCUCCCUGUAUGCCAU    1_0 2                      GCGUAUGAGGAGCCAUGCA     0_4 1
           GCCUGGCUCCCUGUAUGCCAU    1_0 2                      GCGUAUGAGGAGCCAU        0_2 1
          UGCCUGGCUCCCUGUAUGC       0_3 5                      GCGUAUGAGGAGCCAUGCAU    0_2 1
          UGCCUGGCUCCCUGUA          0_7 5                            GAGGAGCCAUGCAUA   1_0 1
          UGCCUGGCUCCCUGU           1_0 6                    GGCGUAUGAGGAGCCAUGCAU    1_0 1
           GCCUGGCUCCCUGUAUGCCA     1_0 3
            CCUGGCUCCCUGUAUGCC      0_1 5


annotation: intergenic
Rpt_annotation: not_repetitive

cal_0.000000_0.000000_root_0.000000_0.000000_leaf_0.000000_0.000000_sUem_0.553398_0.000000


DUPLEX 5:UGCCUGGCUCCCUGUAUGCCA 3:GCGUAUGAGGAGCCAUGCAU vvi-MIR160e     2,1     0.998626
DUPLEX 5:UGCCUGGCUCCCUGUAUGCCA 3:GGCGUAUGAGGAGCCAUGCAU vvi-MIR160e    1,1     0.998626
DUPLEX 5:UGCCUGGCUCCCUGUAUGCC 3:GCGUAUGAGGAGCCAUGCAU vvi-MIR160e      1,1     0.998626
DUPLEX 5:UGCCUGGCUCCCUGUAUGCCA 3:GCGUAUGAGGAGCCAUGCAUA vvi-MIR160e    2,2     0.99847
DUPLEX 5:UGCCUGGCUCCCUGUAUGCC 3:GCGUAUGAGGAGCCAUGCAUA vvi-MIR160e     1,2     0.99847
```

Fig. 4.9 – Output file generated
The output file contains comprehensive information about the predicted locus. Position in 12x grape genome assembly (genomic coordinate, strand and chromosome); sequence, secondary structure and thermodynamic stability; Tags distributions along the sequence (in colour tags that form duplex 2,2 or 2,1 or 1,2 or 1,1; in black sequences un-paired) ; Expression of each tag in one or more tissues (values separated by "_", for example the sequence in red is 254 in leaf and 143 in root); number of time that tag maps into the genome (for the sequence in red is 3). Annotation of the hairpin into the genome (in this case intergenic and not falling in a repetitive region). Information about transcription of precursor in callus, root, leaf, stem from RNAseq experiments (see chapter3).
Further down are shown all pairs of observed reads that form perfect or imperfect miRNA/miRNA*. For example the tag nearest to the 5' (in red) forms with the tag nearest to the 3' (in blue) a duplex in which the 3' of the sequence in red is shifted 2 nt respect to the 5' of the sequence in blue, while the 3' of the sequence in blue is shifted only 1 nucleotide respect to the 5' of the sequence in red (combination 2,1). Because reads can derive from different nested or overlapped hairpins the SVM score is shown for each pair of tags (for the red tag is 0.998626)

4.3 Experimental validation of the bioinformatics pipeline

4.3.1 Datasets

The bioinformatics pipeline described above has been tested using smallRNA deep sequence data from the grapevine, *Vitis vinifera*, generated using standard protocols with the Illumina Genome Analyser platform. In short, 3 lanes of sequence data were obtained from the public website of the Comparative Sequencing of Plant smallRNA project (http://smallrna.udel.edu/data.php). These data derive from leaf, flower and berry samples from the Merlot cultivar. These tissue samples were provided by Gabriele Di Gaspero of the University of Udine (Italy). Additional Libraries for sequencing were constructed in the lab of Prof. Blake Meyers in the University of Delaware (USA) by Emanuele De Paoli and were sequenced by Illumina using standard protocols (http://smallrna.udel.edu/methods.php). For these samples, the adapter sequences were already removed from the publicly available sequence data.

A further two lanes of sequence data were obtained from immature leaf and root tissues from the Pinot Nero genotype pn40024. For these libraries, material was provided by Prof. Mario Pezzotti of the University of Verona (Italy). Libraries were prepared by Erica Mica (Scuola Sant'Anna, Pisa, Italy) according to standard protocols (Mica, Piccolo et al. 2010) and sequencing was performed at the Istituto di Genomica Applicata (Udine, Italy).
The numbers of reads produced and analysed are shown in Tab.4.1. The analytical pipeline was applied independently to the two sets of data although different tissue samples were considered together within single experiments. A total of 19117124 reads that mapped perfectly to the 12x *Vitis vinifera* genome assembly (http://www.plantgdb.org/VvGDB/) were considered.

We observe distinct peaks at 21 and 24 bases in length, as expected for plant smallRNA data, with the 24 base peak expected to correspond predominantly to heterochromatic siRNAs and the 21-22 base peak representing microRNAs, ta-siRNAs and other classes of siRNAs (Tab4.1, Fig.4.10 and Fig4.11). More distinct 24 base map positions are observed (Freq_map_tissue), although the proportion of individual reads mapped tends to be relatively higher for 21 base reads (Freq_read_mapped_tissue) (this is typical of microRNAs that are excised more precisely than heterochromatic siRNAs). These data are in accord with patterns

observed by other workers (Moxon, Jing et al. 2008) suggest that the quality of smallRNA sequence data and mapping is good.

Experiment	total reads	total reads with adapter	total different reads	total mapped reads
Merlot leaf	N/A	3810622	875268	2583036
Merlot flower	N/A	2208760	651490	1511148
Merlot Berry	N/A	2258151	579233	1317172
pn40024 leaf	15413571	10563362	1871362	8368625
pn40024 root	23854642	13951895	1172665	5337143

Tab 4.1 – Number of read produced and analysed
In Fig.4.10(A,B) and Fig.4.11(A,B,C) is shown the length distributions of mapped reads for each tissue sample analysed

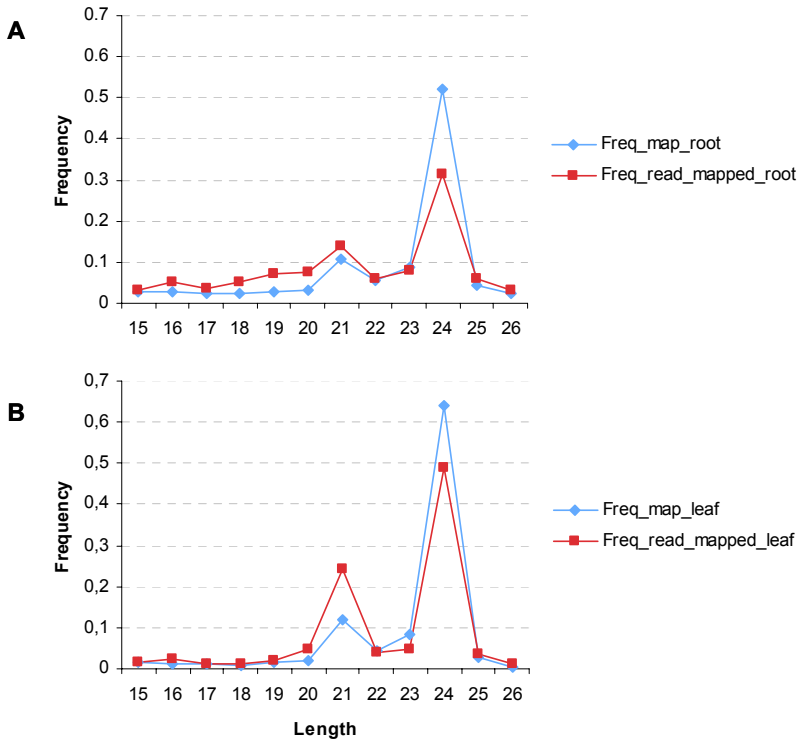


Fig.4.10 - Tissues from pn40024
Length distributions of mapped reads for each tissue sample analysed
A= root, B=leaf

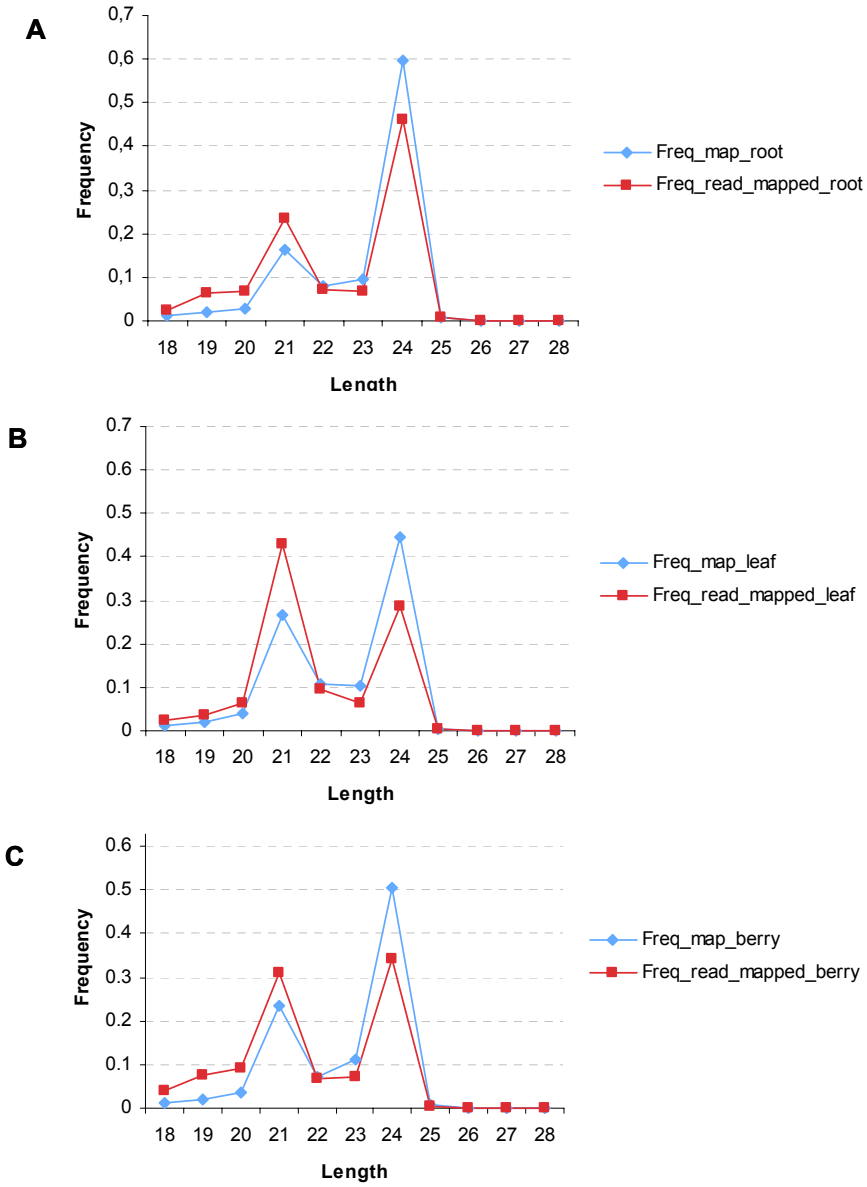


Fig. 4.11 - Tissues from Merlot
Length distributions of mapped reads for each tissue sample analysed
A= root, B=leaf; C=berry

4.4 Preliminary results and identification of conserved miRNA precusors

The independent analyses of the Meyers lab and locally generated smallRNA datasets generated 460 and 1617 loci respectively that passed all preliminary filters. The higher number of candidate loci associated with the locally generated data are likely to correspond at least in part to the much higher number of reads analysed in these experiments (Tab.4.1).

As a first evaluation of how effective our pipeline was, we compared the chromosomal coordinates of our predicted precursors with the coordinates previously annotated as encoding conserved miRNAs detected by microHARVESTER. The Meyer's lab data recovered 57 already known conserved miRNAs, while the locally generated small RNA data allowed the identification of 69 conserved miRNAs.

The discrepancy between these numbers and the 140 predicted conserved precursor loci may be explained by several considerations. First, the pipeline employed here in the analysis of the Illumina data requires that both the miRNA and miRNA* sequences should be represented in the dataset. For miRNAs expressed at low levels, the miRNA* sequence may not be represented in the data and thus the locus is effectively invisible to the method. Secondly, some miRNAs are expressed in strictly tissue or developmental phase specific conditions and may not be expressed in the tissues sampled here. Thirdly, it is not impossible that some small proportion of the miRNA precursors predicted by microHARVESTER (Dezulian, Remmert et al. 2006) are not in fact real/functional miRNA loci. In fact, we previously showed that even attempts to validate precursors predicted through comparative approaches using smallRNA deep sequence data do not allow the validation of all predicted conserved precursors. Accordingly, we conclude that the pipeline implemented here shows an acceptable level of sensitivity in the recovery of precursors for conserved miRNA families.

Here we present several typical outputs of our pipeline (first for several conserved and previously annotated miRNA loci).

Fig.4.12 shows the output for a locus corresponding to a member of the miR160 family. The genomic sequence (oriented 5'-3'), the predicted secondary structure and free energy of folding in kcal/mol as well as the positions and frequencies that reads mapping to the region were observed in two tissue types. The final number associated with smallRNA reads shows the number of genomic loci in which this reads maps perfectly. The next lines show the status of the locus in terms of gene and repeat annotations. The final number associated with smallRNA reads shows the number of genomic loci in which this reads maps perfectly. The next lines show the status of the locus in terms of gene and repeat annotations. The mean coverage for each base in the precursor from RNAseq (Illumina deep sequencing of polyA+ RNA) in four different tissues is displayed. For each tissue, two values are given, the first corresponding to uniquely mapping RNAseq reads, the second for reads that map in redundant genomic loci. Finally, pairs of smallRNA reads that were initially used to define the hairpin (from the miRNA/miRNA* filter) are shown along with their degree of overhang

(2,2 means the canonical 2 base 3' overhangs are present). the final value are SVM probability scores for the locus representing a real miRNA. It is notable that the "correct" 2,2 miRNA/miRNA* pairings correspond to the two most frequently observed reads.

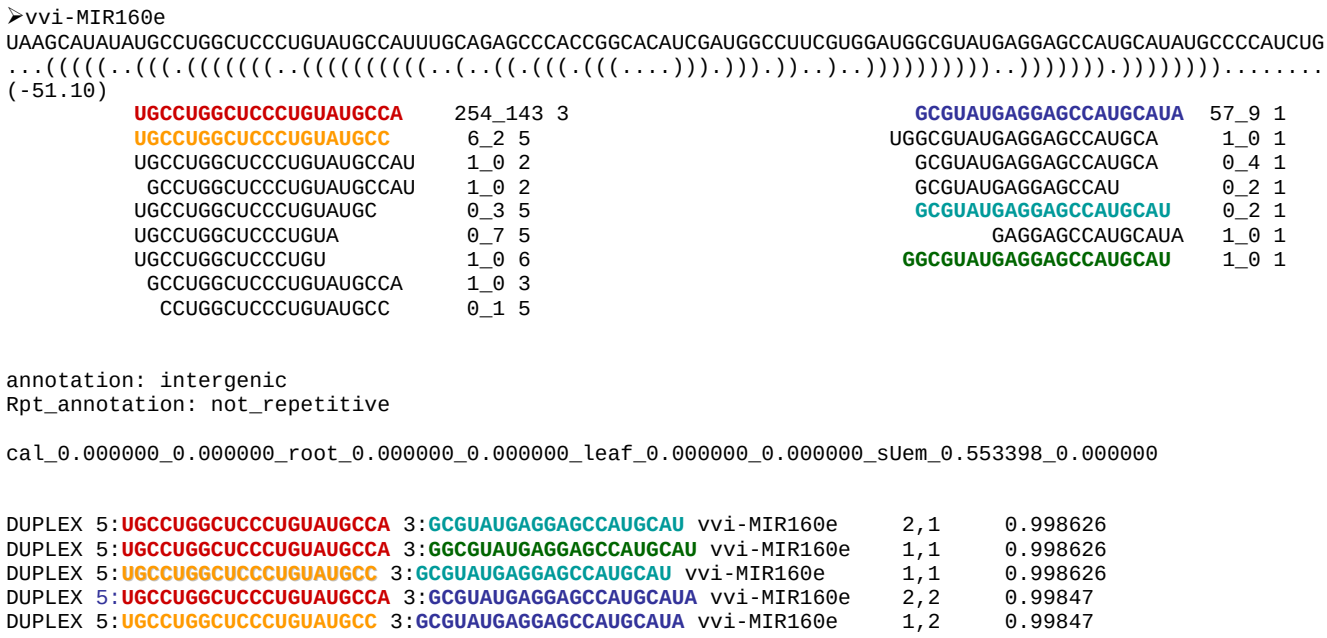


Fig.4.12 - A typical output of our pipeline corresponding to the conserved miRNA vvi-MIR160a from pn40024 set of data
The most expressed read is shown in red (254 in leaf, 143 in root). Its correspondent miRNA* could be the blue sequence that is the second most expressed tag (57 in leaf, 9 in root). Interestingly these two tag form a perfect duplex (2.2) and in particular the 'putative miRNA*' maps only one time into te genome.

Reads are concentrated in two regions of the hairpin and corresponds to the real mature miRNA and its miRNA*. Slightly shifted reads observed with low frequency and correspond to imprecision in DCL-1 cutting, a typical observation for miRNA sequences. No reads from the antisense strand were observed in this case. The most expressed tags corresponds to mature miRNA and miRNA*. This prediction derived from the locally generated data and we can see that the mature miRNA was observed 254 times in leaf and 143 times in the root sample, it maps to 3 distinct genomic loci (2 others apart from this one). Its corresponding miRNA* was observed 57 times in leaf and 9 times in the root sample and maps uniquely to this chromosomal location. Taken together, these observations suggest that this particular miR160 locus is expressed in our tissue samples. The colored sequences correspond to reads involved in a perfect miRNA/miRNA* pairings (2,2) or not (1,2 ; 1,1; 2,1).

This miRNA was originally discovered in cotton, *Gossypium hirsutum* (Pang, Woodward et al. 2009) and was recently annotated also in the grapevine genome (Pantaleo, Szittya et al. 2010). We observe highest expression of the smallRNA deriving from the 5' arm of the hairpin, while the cotton annotation suggest that the 3' smallRNA should be the mature miRNA. Wwe performed a target prediction analysis for both tags of the duplex (UUCCAUCUCUUGCACACUGGA in red and UGGUGUGCACGGGAUGGAAUA in blue) using the Axtell target finder script from the CLEAVELAND package (Addo-Quaye, Miller et al. 2009). We found a probable target for only for the sequence most expressed UUCCAUCUCUUGCACACUGGA in our dataset (in red in Fig.4.12). The target is the *Vitis* ortholog of petunia DOUBLE TOP, an F-Box gene related to *Arabidopsis* UNUSUAL FLORAL ORGANS (UFO) which is involved in maintenance of meristematic identity during floral development(Souer, Rebocho et al. 2008). Interestingly, we see this miRNA expressed most highly in floral tissues, potentially indicating a role for miR2950 in fine regulation of F-Box gene expression during floral development.

Fig.4.13 shows the vvi-miR2950 locus as represented in the Meyers dataset.

```
>vvi-MIR2950
CUUGUGAUGUAUUCCAUCUCUUGCACACUGGACCAGCGCUCCAGCUGCAGUUUGGUGUGCACGGGAUGGAAUACAUCAUGGAUUC
.(..((((((((((((((((.(((((((..(((((((......)))))..)))..)))))))).)))))))))))))))..)....   (-
52.60)
           UUCCAUCUCUUGCACACUGGA     10_278_54 1       UGGUGUGCACGGGAUGGAAUA      10_108_20 1
          AUUCCAUCUCUUGCACACUGG       0_15_5 1         UGGUGUGCACGGGAUGGAAUAC      0_1_0 1
           UUCCAUCUCUUGCACACUGG       2_10_0 1         UGGUGUGCACGGGAUGGAAU        0_1_0 1
           UUCCAUCUCUUGCACACUGGA      0_4_0 1           GGUGUGCACGGGAUGGAAUAC      0_2_0 1
           UUCCAUCUCUUGCACACUGGACCA   1_3_0 1
          AUUCCAUCUCUUGCACACUGGACC    0_1_0 1
          AUUCCAUCUCUUGCACACUGGA      0_1_0 1
             CCAUCUCUUGCACACUGGACC    0_1_0 1
             CCAUCUCUUGCACACUGGAC     0_0_1 1
              CAUCUCUUGCACACUGGA           0_2_0 1


annotation: intergenic
Rpt_annotation: not_repetitive

cal_0.552941_0.000000_root_0.000000_0.000000_leaf_0.000000_0.000000_stem_0.164706_0.000000


DUPLEX 5:AUUCCAUCUCUUGCACACUGGA 3:UGGUGUGCACGGGAUGGAAUA    2,1     0.998257
DUPLEX 5:AUUCCAUCUCUUGCACACUGG  3:UGGUGUGCACGGGAUGGAAUA    1,1     0.998257
DUPLEX 5:AUUCCAUCUCUUGCACACUGG  3:GGUGUGCACGGGAUGGAAUAC    2,2     0.998182
DUPLEX 5:AUUCCAUCUCUUGCACACUGG  3:UGGUGUGCACGGGAUGGAAUAC   1,2     0.998182
DUPLEX 5:UUCCAUCUCUUGCACACUGGA  3:UGGUGUGCACGGGAUGGAAU     2,1     0.998313
DUPLEX 5:UUCCAUCUCUUGCACACUGG   3:UGGUGUGCACGGGAUGGAAU     1,1     0.998313
DUPLEX 5:UUCCAUCUCUUGCACACUGGA  3:UGGUGUGCACGGGAUGGAAUA    2,2     0.998316
DUPLEX 5:UUCCAUCUCUUGCACACUGG   3:UGGUGUGCACGGGAUGGAAUA    1,2     0.998316
```

Fig.4.13 Output for MiR2950 from Merlot set of data
The most expressed tag corresponds to the red (10 in leaf, 278 in root, 54 in flower). Its correspondent miRNA* is the blue sequence that is the second most expressed tag (10 in leaf, 108 in root, 20 in flower). Interestingly these two tag form a perfect duplex (2.2) and are both unique mapping tags.

As with the previous example, the most frequenct reads correspond to the canonical miRNA/miRNA* pair, the locus falls in an intergenic, non-repetitive region and the Support Vector Machine analysis strongly supports the locus as a valid pre-miRNA.

4.5 Novel *and lineage specific* miRNA precursors in the grapevine, *Vitis vinifera*

Having confirmed the capacity of our method to recover conserved, already annotated miRNAs, such candidates were discarded. Careful manual consideration of remaining candidates allowed the identification of over 80 potentially grape-specific miRNAs that follow the standard rules for miRNA annotation. Most of these (around 50) were identified in both the locally generated and Meyers datasets. Around 10 loci consistently produce 24 nt rather than 21 nt mature miRNAs. This finding was unexpected because the DCL-1 protein usually processes miRNA into 21mers. In addition we found that *lineage specific* miRNAs tend to be longer with respect to conserved miRNAs. Many of them are more or less 500 nt long.

Interestingly we detected that many of these *lineage specific* loci produce large quantities of more than 1 mature miRNA.

The output of our pipeline for several examples of candidate novel miRNA loci with interesting characteristics are presented in the followings paragraphs,

4.6 24 base miRNAs

Fig.4.14 shows an example of a *lineage specific* miRNA recovered from the locally generated data. As with the previosu examples, reads are concentrated in discrete positions on opposite arms of the hairpin, and the most frequently observed reads correspond to a canonical miRNA/miRNA* pair and the hairpin yields a good Support Vector Machine Score. Strikingly, the candidate miRNA and its corresponding miRNA* are both 24 bases long. It is also of note that the precursor falls within an annotated intron of a RAB type GTP binding protein gene. Manual checks confirmed the intronic status and target prediction suggested that the miRNA could target a CAAX amino terminal protease family protein (involved in maturation and membrane targeting of RAS/RHO/RAB proteins). Thus, the bioinformatics analysis suggests an intruiging naegative feedback regulatory loop between the RAB gene

hosting the miRNA and the target of the miRNA. We might speculate that in physiological conditions, an excess of the RAB transcript could lead to an increase of miRNA production and concordant reduction of levels of the CAAX protease (in turn reducing the levels of correctly matured RAB protein). To our knowledge, this miRNA represents one of the first strongly supported instances of intronic miRNAs in plants and experimental studies of the functional role and regulatory mechanisms of the miRNA will be of great interest.

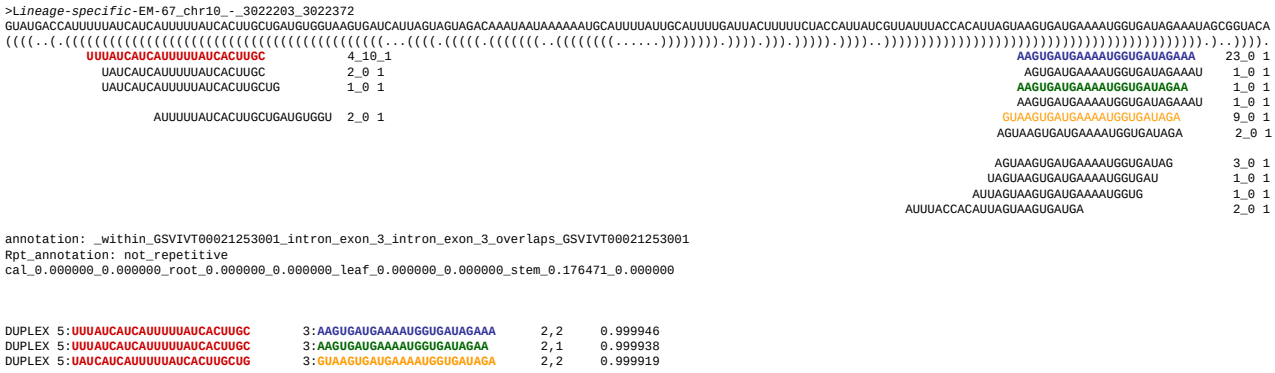


Fig. 4.14 Output for a novel miRNA from pn40024 set of data
The most expressed tag corresponds to the blue (23 in leaf, 0 in root). Its correspondent miRNA* is the red sequence that is the second most expressed tag (4 in leaf, 10 in root). Interestingly these two tag form a perfect duplex (2.2) and are unique mapping tags

4.7 Conserved atypical processing of miRNA precursors

The classical model of miRNA maturation is that a first DCL1 cut occurs about 10-11 bases from the bottom of the stem, and the second cut 21 bases further up the stem (Bernstein, Caudy et al. 2001; Kurihara and Watanabe 2004; Vermeulen, Behlen et al. 2005). This model of action is supported by detailed molecular studies in several *Arabidopsis* miRNAs (Mateos, Bologna et al. 2010).

Recently some miRNAs (miR159 and miR319) were shown to be first processed near the loop, and several 'phased' cuts occur down the stem (Fig.4.15). Mutations and deletions in the upper part of the stem region were shown to abolish processing of these precursors in *Arabidopsis* (Bologna, Mateos et al. 2009; Schwab and Voinnet 2009) while such changes had no effect on the processing of other miRNAs (Mateos, Bologna et al. 2010).

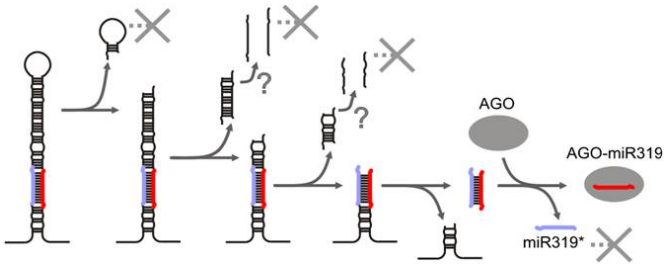


Fig.4.15 - Model of the maturation of *Arabidopsis* miR319 and 159
From (Bologna, Mateos et al. 2009)

As consequence of these sequential cuts, the processing is "phased" and several small RNAs, are consequentially produced and these might be detected through NGS data analysis. In *Arabidopsis*, this alternative way of processing was experimentally validated for miR159 and miR319 by the 5' RACE–PCR strategy (Bologna, Mateos et al. 2009) Bologna et al. (2009) also generated some stem loops mutants for miR319 and miR159 precursors and noted that deletions of part of the stem nearest of the loop inhibited miR159 expression (Fig.4.16).

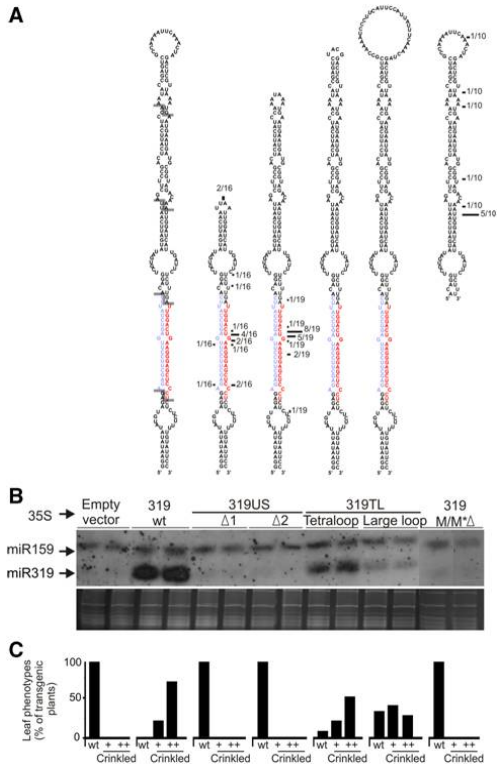


Fig.4.16 - Sequence determinants for miR319a processing.
(A) Scheme showing the stem loops of several mutant miR319 precursors. The cleavage sites analysed by 5′ RACE method are indicated by black lines on the right of each precursor when determined. 319 wt corresponds to 319LS2. The miRNA is indicated in red and the miRNA* in blue. (B) Small-RNA blots showing accumulation of miR319. Two pools of 25 independent transgenic plants expressing the corresponding precursor from the 35S promoter were analysed in each case. (C) Analysis of leaf shape in at least 100 independent transgenic plants for each construct. From (Bologna, Mateos et al. 2009)

To investigate whether this mechanism is conserved in *Vitis vinifera*, we carefully considered the distribution of reads mapping to annotated miR159 loci. Indeed, the distribution of reads on the miR159c precursor strongly suggests that this mechanism is conserved (Fig.4.17).

The read distribution in vvi-miR159c suggests that three cuts are performed towards the loop with respect to the mature miR159. Given the experimental data for *Arabidopsis*, it is highly probable that the cuts are also performed from the loop towards the base also in *Vitis*. In fact in vvi-miR159c, the most expressed pair of tags maps into the base of the stem and correspond to the expected mature miRNA (Fig.4.17).

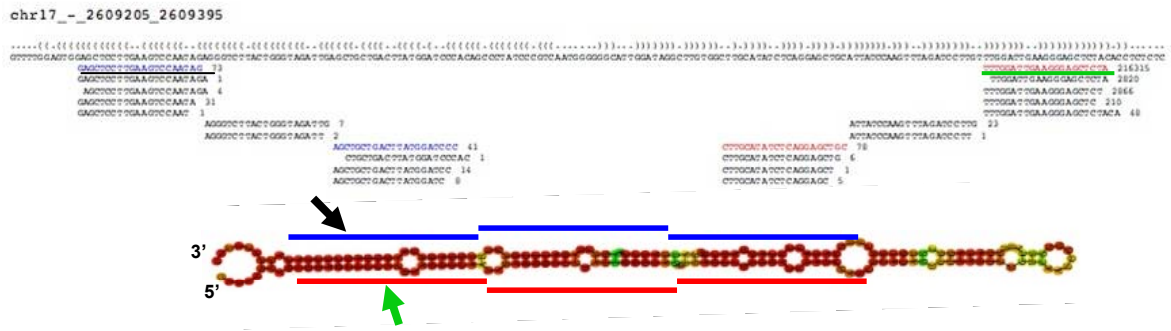


Fig. 4.17 - vvi-miR159c
The putative mature miRNA is the most expressed tag (it is highlighted in green in the alignment and indicated with the green arrow in the secondary structure). The correspondent miRNA* is the second most expressed tag (highlighted in black in the alignment and indicated with the black arrow in the secondary structure). The tags have a phased distribution

The same phased read distribution was found in vvi-miR169x (Fig.4.18), suggesting that a similar mechanism of processing could not be excluded (although we do not have additional experimental support in grape or other plants). Without experimental validations, we are not able to define the direction of the cleavage series (from the beginning of the stem or from the loop), but, our data show that, as in vvi-miR159c, the most expressed tags map into the extreme part of the stem (as usually happens for functional miRNAs) and that the distribution of tags is phased. Thus, the "top-down" phased production of miRNAs is likely to be present in additional deeply conserved miRNA families.

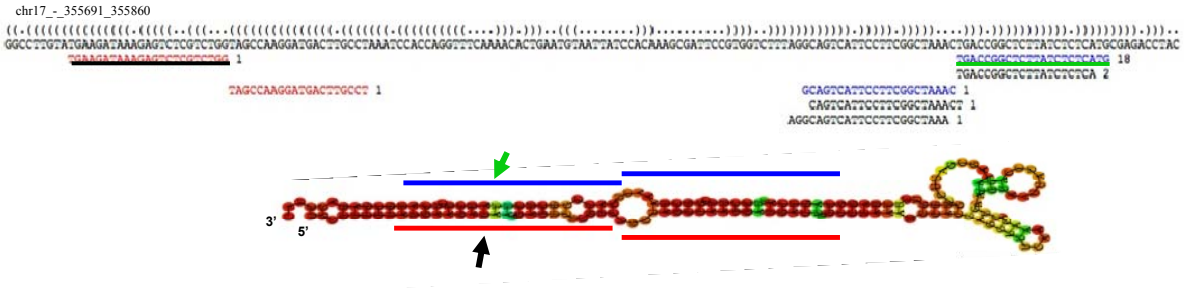


Fig.4.18 - vvi-miR169x
The putative mature miRNA is the most expressed tag (it is highlighted in green in the alignment and indicated with the green arrow in the secondary structure). The correspondent miRNA* is the second most expressed tag (is highlighted in black in the alignment and indicated with the black arrow in the secondary structure). The tags seems to have a phased distribution.

4.8 Phased smallRNA production from *lineage specific* miRNAs

Our pipeline also recovered several cases of novel pre-miRNAs that present notable phasing of smallRNA products. Fig.4.19, illustrates a novel, apparently non-conserved locus showing a similar pattern of reads as observed previously for the conserved miRNAs miR159 / 319 / 169. Because of the length of its secondary structure, part of this had been cut in nucleotide sequence., while is shown in the entire structure.

We detected that the most expressed tags map into the extreme basal part of the 5' stem.

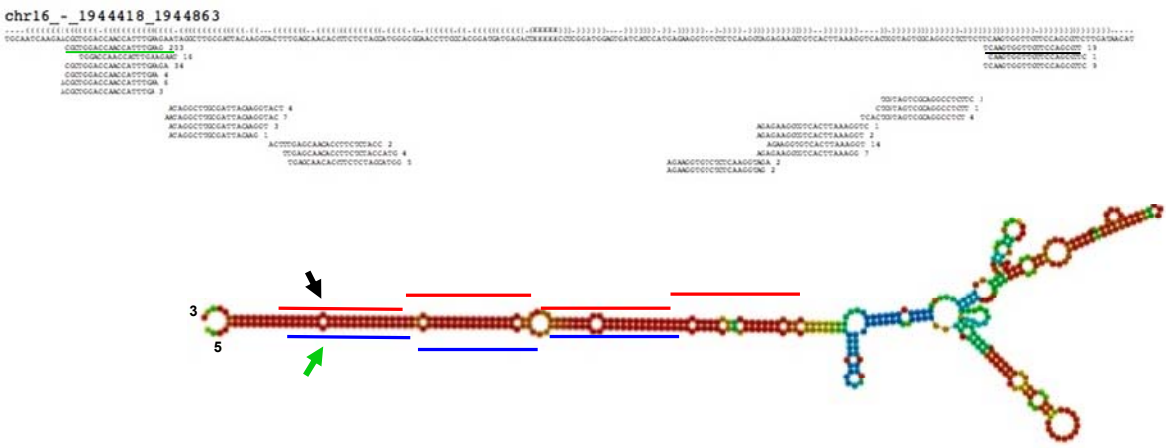


Fig.4.19 - Novel miRNA
The putative mature miRNA is the most expressed tag (it is highlighted in green in the alignment and is indicated with the green arrow in the secondary structure). The correspondent miRNA* is the second most expressed tag (is highlighted in black in the alignment and indicated with the black arrow in the secondary structure). The tags seems to have a phased distribution

Our data therefore suggest that the phased production of miRNA-like molecules from hairpin structures may be more prevalent than previously appreciated, both for conserved and *lineage specific* miRNAs.

4.9 Implications for the evolution of miRNA precursors.

A recent model for the origin and evolution of miRNA genes (Vazquez, Legrand et al. 2010) (Fig.4.20) suggests that that pre-miRNAs could derive from transcribed inverted repeats.

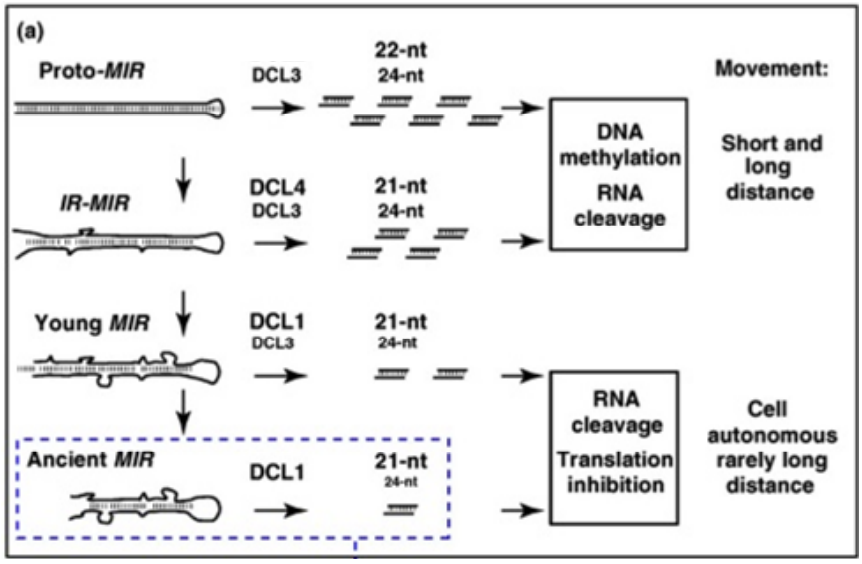


Fig.4.20 – Model of Vazquez et al (Vazquez, Legrand et al. 2010): the microRNA and microRNA-like pathways in an evolutionary perspective.

A novel inverted repeat would be expected to form a perfectly complementary hairpin, which might be expected to represent a good substrate for the enzyme DCL3, which is also involved in the production of heterochromatic siRNAs and which tends to generate imprecisely excised 24 base RNAs. Such a template might also be recognized by DCL2 which is also involved in cleavage of viral dsRNAs and which produces 22 base RNAs. Under the model of Vazquez, substitutions and small insertions and deletions could, over the course of time, lead to the evolution of hairpins with more mismatches and bulges. Such hairpins would be expected to represent better targets for DCL4 (which is also involved in the production of phased smallRNAs in the ta-siRNA pathway) and could lead to more specific production of 21 base RNAs, potentially though a phased mechanism of smallRNA production. Additional mutations would tend to lead to shorter stem-loop structures, which, in the presence of

selective pressures favouring production of specific smallRNAs would result in the stabilization of good DCL1 templates and the evolution of canonical pre-miRNAs.

Thus, during the evolution of a typical miRNA gene, different DCL enzymes might be responsible for processing. Conversely, this model might be used to generate several testable hypotheses:

- *lineage specific* (young) pre-miRNAs might tend to possess longer and more complementary stems than ancient conserved miRNAs;
- production of 24 base miRNA-like molecules might be more common in *lineage specific* miRNAs than in ancient, conserved miRNAs;
- production of phased smallRNAs from hairpin structures might be more common in younger miRNAs.

Evolution of MIR genes by progressive random mutations in initially perfect inverted repeats (IRs) which progressively yield shorter hairpins with more mismatches and bulges. Evolution of IR genes into MIR genes is accompanied by a change in hairpin processing by DCLs and in the size of the miRNAs/miRNA-like siRNAs generated. Long IR use primarily DCL2, recently-evolved MIR with intermediate-sized hairpins use DCL4 and ancient MIR genes with short hairpins use primarily DCL1 to produce miRNAs. Moreover, all hairpins are processed to different extents by DCL3 to yield long-miRNAs. The shift in DCL usage during MIR gene evolution might also be accompanied by changes in the function of the miRNAs/miRNAs-like siRNAs generated (Vazquez 2006)
In general, the predictions of this model are consistent with observations from our data and bioinformatics pipeline. We do recover many long precursors among our *lineage specific* candidate miRNAs and we observe loci producing 24 base miRNA-like molecules among the same set. Additionally, we observe many loci that generate strand specific phased smallRNAs from hairpin precursors.
Fig.4.21 shows a novel pre-miRNA-like locus that we find particularly interesting in the context of this evolutionary model.

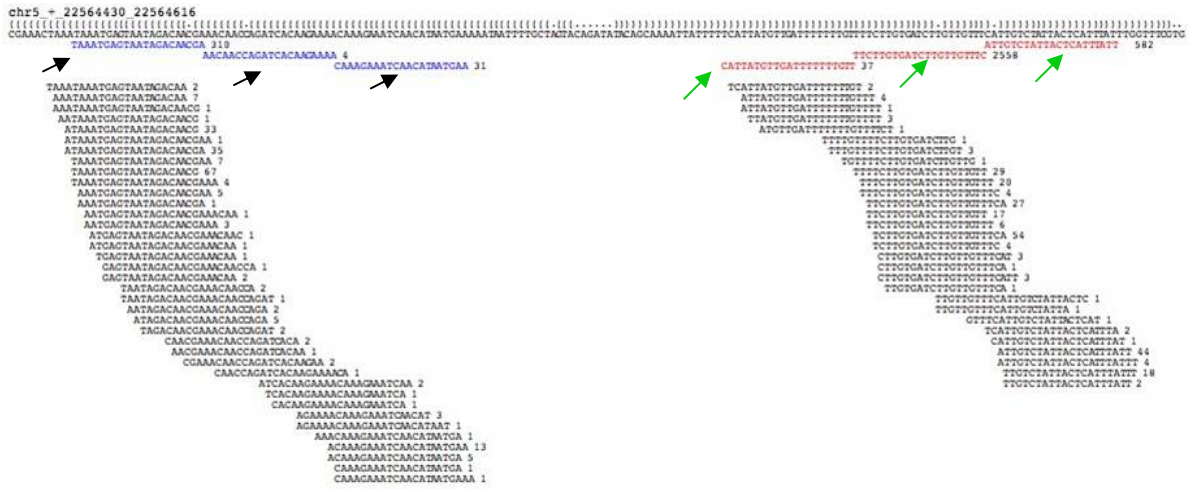


Fig.4.21 - A particular case of putative-novel miRNA
The most expressed tags map into the extreme basal part of the 3'

The hairpin is relatively long and presents few mismatches and bulges in its structure, consistent with it being a relatively young inverted repeat. The precursor shows no significant similarity with pre-miRNAs in miRBase. Reads are derived not from discrete locations on either arm of the hairpin but are apparently distributed along both arms. However, the most frequently observed reads (in red and blue) demonstrate clear patterns of phasing as seen in miR159 and other examples presented. While the majority of reads observed are 21 bases in length, we are tempted to view this locus as representing a transitional state between early phases of the evolutionary model shown above (a mixture of aspecific and phased smallRNA production). We have no information as to whether this locus plays a physiological role in the grapevine.

It is interesting that several relatively ancient miRNA loci (miR159, miR319 and miR169 are all present in both monocots and dicots) produce phased smallRNAs as, under the evolutionary model described, phased miRNA production should be considered a typical marker of younger miRNAs. However, it is easy to imagine an explanation for this observation. If during the evolution of a miRNA gene, the production of a particular smallRNA is beneficial for the organism, purifying selection will tend to maintain its sequence and expression. If such a smallRNA is produced through a phasing mechanism, and if it falls towards the base of a stem structure, it is difficult for substitution events to shorten the hairpin while maintaining the sequence of the smallRNA. If the model proposed is valid, one would expect only loci where the functional smallRNAs derives from near the loop to progress towards shorter hairpins and processing exclusively mediated by DCL1.

4.10 Conclusions

We have implemented a pipeline for the identification of plant miRNA precursors from smallRNA Next Generation Sequence data. Our method uses novel heuristics to exclude spurious candidates and select potential pre-miRNA loci. We have incorporated the Support Vector Machine classifier presented in chapter 3 as an additional filter, and our pipeline links with annotations of genes and repeats as well as transcriptome data to furnish additional information on the loci recovered.

We show that the method is capable of recovering annotated miRNAs from *Vitis*, without use of sequence similarity to identify candidates. The method recovers many novel, canonical miRNAs from *Vitis* and is capable of identifying loci producing miRNA-like smallRNAs with characteristics that are atypical of most conserved miRNAs, for example 24 base miRNAs and loci producing multiple phased hairpin-derived smallRNAs. It is worth noting that "classical" tools for the identification of miRNAs from high throughput sequence data are not designed to recover phased miRNAs. MiRDeep (Friedlander, Chen et al. 2008) for example searches for short peaks of read density that should be shorter than 30bases in length. Loci producing high quantities of phased reads will not conform to this expectation and are unlikely to be detected by such software.

The patterns of smallRNA generated from putatively *lineage specific* loci have been considered in the context of a current model of miRNA gene evolution, and we find broad agreement with the expectations of this model. Accordingly, we believe that, as well as its application in discovery of novel miRNAs, our approach will find application in the study of miRNA gene evolution and the further development of such models.

While the system implemented here has been designed with plant miRNAs in mind, there is no obvious reason why it should not be used to study miRNAs in animal or other systems. Indeed, preliminary data not presented here suggest that even without modification of parameters imposed for plant miRNAs, the pipeline functions well for animal deep sequence data (over 250 known human miRNAs were recovered using around 25 million publically available human smallRNA reads).

Here we have used the Support Vector Machine classifier based on exclusively hairpin derived properties to prioritize predictions that passed all filters in the miRNA prediction pipeline. However, deep sequencing of smallRNAs of course also offers the opportunity to identify candidate mature miRNAs. One future development could be to incorporate additional features based on the structural and sequence characteristics of the candidate

miRNA/miRNA* regions into the SVM in order to provide a more comprehensive set of features to differentiate real pre-miRNAs from non-miRNA hairpins. It is envisaged that such a strategy should increase both the sensitivity and specificity of the classification system.

Chapter 5

5.1 General discussion

This thesis presents the development, implementation and testing of bioinformatics strategies for comparative, *ab-initio* and deep-sequencing based miRNA discovery. Comparative predictions were validated through deep sequencing and differential expression in *Vitis* tissues demonstrated through oligonucleotide array experiments. Additionally, novel bioinformatics strategies for the definition of primary miRNA transcript coordinates and splicing patterns from whole transcriptome sequencing data were implemented. A novel support vector machine-based *ab-initio* miRNA prediction software was implemented and tested. The software outperforms similar published approaches, particularly in terms of the low rate of false positive predictions generated. A new strategy for the identification of miRNAs and their precursors from smallRNA Next Generation Sequence data was implemented and tested extensively in the grapevine, *Vitis vinifera*. The method performs well in the recovery of known miRNAs and identifies many high confidence predictions for novel or *lineage specific* miRNAs. Interestingly, our method is also able to identify many loci that resemble miRNAs but with unusual patterns of processing. The patterns of smallRNA generation at such loci fit current models of miRNA gene evolution extremely well and lead us to believe that many of these loci represent "transitional forms" in the origin of miRNA genes. Thus, we believe that the method described could be of great value in the study of miRNA gene evolution and in the generation of novel hypotheses for mechanisms of miRNA gene origin, selection and evolution.

The work presented in this thesis therefore constitutes a series of software tools and strategies that can be applied generally in plants and to a large extent also in animal species for the detection and annotation of conserved and novel miRNAs.

Bibliography

Abdel-Ghany, S. E. and M. Pilon (2008). "MicroRNA-mediated systemic down-regulation of copper protein expression in response to low copper availability in Arabidopsis." J Biol Chem **283**(23): 15932-15945.

Addo-Quaye, C., T. W. Eshoo, et al. (2008). "Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome." Curr Biol **18**(10): 758-762.

Addo-Quaye, C., W. Miller, et al. (2009). "CleaveLand: a pipeline for using degradome data to find cleaved small RNA targets." Bioinformatics **25**(1): 130-131.

Agarwal, S., C. Vaz, et al. (2010). "Prediction of novel precursor miRNAs using a context-sensitive hidden Markov model (CSHMM)." BMC Bioinformatics **11 Suppl 1**: S29.

Allen, E. and M. D. Howell (2010). "miRNAs in the biogenesis of trans-acting siRNAs in higher plants." Semin Cell Dev Biol **21**(8): 798-804.

Allen, E., Z. Xie, et al. (2005). "microRNA-directed phasing during trans-acting siRNA biogenesis in plants." Cell **121**(2): 207-221.

Allen, E., Z. X. Xie, et al. (2005). "microRNA-directed phasing during trans-acting siRNA biogenesis in plants." Cell **121**(2): 207-221.

Altschul, S. F., W. Gish, et al. (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Altuvia, Y., P. Landgraf, et al. (2005). "Clustering and conservation patterns of human microRNAs." Nucleic Acids Res **33**(8): 2697-2706.

Ambros, V. (2003). "MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing." Cell **113**(6): 673-676.

Ambros, V., B. Bartel, et al. (2003). "A uniform system for microRNA annotation." RNA **9**(3): 277-279.

Amrani, N., M. S. Sachs, et al. (2006). "Early nonsense: mRNA decay solves a translational problem." Nat Rev Mol Cell Biol **7**(6): 415-425.

Andronescu, M. S., C. Pop, et al. (2010). "Improved free energy parameters for RNA pseudoknotted secondary structure prediction." RNA **16**(1): 26-42.

Aravin, A. A., M. Lagos-Quintana, et al. (2003). "The small RNA profile during Drosophila melanogaster development." Dev Cell **5**(2): 337-350.

Arazi, T., M. Talmor-Neiman, et al. (2005). "Cloning and characterization of micro-RNAs from moss." Plant J **43**(6): 837-848.

Aukerman, M. J. and H. Sakai (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, M. J. and D. P. Bartel (2005). "Antiquity of microRNAs and their targets in land plants." Plant Cell **17**(6): 1658-1673.

Bachellerie, J. P., J. Cavaille, et al. (2002). "The expanding snoRNA world." Biochimie **84**(8): 775-790.

Bandres, E., E. Cubedo, et al. (2006). "Identification by Real-time PCR of 13 mature microRNAs differentially expressed in colorectal cancer and non-tumoral tissues." Mol Cancer **5**: 29.

Barad, O., E. Meiri, et al. (2004). "MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues." Genome Res **14**(12): 2486-2494.

Barik, S. (2008). "An intronic microRNA silences genes that are functionally antagonistic to its host gene." Nucleic Acids Res **36**(16): 5232-5241.

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell **116**(2): 281-297.
Bashirullah, A., A. E. Pasquinelli, et al. (2003). "Coordinate regulation of small temporal RNAs at the onset of Drosophila metamorphosis." Dev Biol **259**(1): 1-8.
Batuwita, R. and V. Palade (2009). "microPred: effective classification of pre-miRNAs for human miRNA gene prediction." Bioinformatics **25**(8): 989-995.
Bentley, D. R., S. Balasubramanian, et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-59.
Bentwich, I., A. Avniel, et al. (2005). "Identification of hundreds of conserved and nonconserved human microRNAs." Nat Genet **37**(7): 766-770.
Berezikov, E., V. Guryev, et al. (2005). "Phylogenetic shadowing and computational identification of human microRNA genes." Cell **120**(1): 21-24.
Berezikov, E., F. Thuemmler, et al. (2006). "Diversity of microRNAs in human and chimpanzee brain." Nat Genet **38**(12): 1375-1377.
Bernstein, E., A. A. Caudy, et al. (2001). "Role for a bidentate ribonuclease in the initiation step of RNA interference." Nature **409**(6818): 363-366.
Blaszczyk, J., J. E. Tropea, et al. (2001). "Crystallographic and modeling studies of RNase III suggest a mechanism for double-stranded RNA cleavage." Structure **9**(12): 1225-1236.
Bohnsack, M. T., K. Czaplinski, et al. (2004). "Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs." RNA **10**(2): 185-191.
Bologna, N. G., J. L. Mateos, et al. (2009). "A loop-to-base processing mechanism underlies the biogenesis of plant microRNAs miR319 and miR159." EMBO J **28**(23): 3646-3656.
Bolstad, B. M., R. A. Irizarry, et al. (2003). "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias." Bioinformatics **19**(2): 185-193.
Bonnet, E., J. Wuyts, et al. (2004). "Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences." Bioinformatics **20**(17): 2911-2917.
Borchert, G. M., W. Lanier, et al. (2006). "RNA polymerase III transcribes human microRNAs." Nat Struct Mol Biol **13**(12): 1097-1101.
Borenstein, E. and E. Ruppin (2006). "Direct evolution of genetic robustness in microRNA." Proc Natl Acad Sci U S A **103**(17): 6593-6598.
Borsani, O., J. Zhu, et al. (2005). "Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis." Cell **123**(7): 1279-1291.
Brameier, M., A. Herwig, et al. (2010). "Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs." Nucleic Acids Res.
Brameier, M., A. Krings, et al. (2007). "NucPred--predicting nuclear localization of proteins." Bioinformatics **23**(9): 1159-1160.
Brameier, M. and C. Wiuf (2007). "Ab initio identification of human microRNAs based on structure motifs." BMC Bioinformatics **8**: 478.
Brennecke, J. and S. M. Cohen (2003). "Towards a complete description of the microRNA complement of animal genomes." Genome Biol **4**(9): 228.
Brenner, S., M. Johnson, et al. (2000). "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays." Nat Biotechnol **18**(6): 630-634.
Calin, G. A., C. D. Dumitru, et al. (2002). "Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia." Proc Natl Acad Sci U S A **99**(24): 15524-15529.

Carrington, J. C. and V. Ambros (2003). "Role of microRNAs in plant and animal development." Science **301**(5631): 336-338.
Cerutti, L., N. Mian, et al. (2000). "Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain." Trends Biochem Sci **25**(10): 481-482.
Chen, C. Z., L. Li, et al. (2004). "MicroRNAs modulate hematopoietic lineage differentiation." Science **303**(5654): 83-86.
Chen, X. (2005). "MicroRNA biogenesis and function in plants." FEBS Lett **579**(26): 5923-5931.
Chendrimada, T. P., R. I. Gregory, et al. (2005). "TRBP recruits the DICER complex to Ago2 for microRNA processing and gene silencing." Nature **436**(7051): 740-744.
Chih-Chung Chang and Chih-Jen Lin (2001) LIBSVM : a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
Choi, D., J. H. Kim, et al. (2004). "Whole genome analysis of the OsGRF gene family encoding plant-specific putative transcription activators in rice (Oryza sativa L.)." Plant Cell Physiol **45**(7): 897-904.
Clote, P., F. Ferre, et al. (2005). "Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency." Rna-a Publication of the Rna Society **11**(5): 578-591.
Collier, S., A. Pendle, et al. (2006). "A distant coilin homologue is required for the formation of cajal bodies in Arabidopsis." Mol Biol Cell **17**(7): 2942-2951.
Cortes C, Vapnik V. (1995). Support-vector network. Machine Learning. 20:273–297.
Creighton, C. J., J. G. Reid, et al. (2009). "Expression profiling of microRNAs by deep sequencing." Brief Bioinform **10**(5): 490-497.
Davison, T. S., C. D. Johnson, et al. (2006). "Analyzing micro-RNA expression using microarrays." Methods Enzymol **411**: 14-34.
Denoeud, F., J. M. Aury, et al. (2008). "Annotating genomes with massive-scale RNA sequencing." Genome Biol **9**(12): R175.
Dezulian, T., M. Remmert, et al. (2006). "Identification of plant microRNA homologs." Bioinformatics **22**(3): 359-360.
Dong, Z., M. H. Han, et al. (2008). "The RNA-binding proteins HYL1 and SE promote accurate in vitro processing of pri-miRNA by DCL1." Proc Natl Acad Sci U S A **105**(29): 9970-9975.
Dostie, J., Z. Mourelatos, et al. (2003). "Numerous microRNPs in neuronal cells containing novel microRNAs." RNA **9**(2): 180-186.
Eamens, A. L., N. A. Smith, et al. (2009). "The Arabidopsis thaliana double-stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes." RNA **15**(12): 2219-2235.
Elbashir, S. M., J. Harborth, et al. (2001). "Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells." Nature **411**(6836): 494-498.
Elbashir, S. M., W. Lendeckel, et al. (2001). "RNA interference is mediated by 21- and 22-nucleotide RNAs." Genes Dev **15**(2): 188-200.
Emery, J. F., S. K. Floyd, et al. (2003). "Radial patterning of Arabidopsis shoots by class III HD-ZIP and KANADI genes." Curr Biol **13**(20): 1768-1774.
Ender, C., A. Krek, et al. (2008). "A human snoRNA with microRNA-like functions." Mol Cell **32**(4): 519-528.
Fang, Y. and D. L. Spector (2007). "Identification of nuclear dicing bodies containing proteins for microRNA biogenesis in living Arabidopsis plants." Curr Biol **17**(9): 818-823.

Forstemann, K., Y. Tomari, et al. (2005). "Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein." PLoS Biol **3**(7): e236.
Freier, S. M., R. Kierzek, et al. (1986). "Improved free-energy parameters for predictions of RNA duplex stability." Proc Natl Acad Sci U S A **83**(24): 9373-9377.
Friedlander, M. R., W. Chen, et al. (2008). "Discovering microRNAs from deep sequencing data using miRDeep." Nat Biotechnol **26**(4): 407-415.
Gauwerky, C. E., K. Huebner, et al. (1989). "Activation of MYC in a masked t(8;17) translocation results in an aggressive B-cell leukemia." Proc Natl Acad Sci U S A **86**(22): 8867-8871.
German, M. A., M. Pillay, et al. (2008). "Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends." Nat Biotechnol **26**(8): 941-946.
Golan, D., C. Levy, et al. (2010). "Biased hosting of intronic microRNA genes." Bioinformatics **26**(8): 992-995.
Golden, T. A., S. E. Schauer, et al. (2002). "SHORT INTEGUMENTS1/SUSPENSOR1/CARPEL FACTORY, a DICER homolog, is a maternal effect gene required for embryo development in Arabidopsis." Plant Physiol **130**(2): 808-822.
Gordon, L., A. Y. Chervonenkis, et al. (2003). "Sequence alignment kernel for recognition of promoter regions." Bioinformatics **19**(15): 1964-1971.
Grad, Y., J. Aach, et al. (2003). "Computational and experimental identification of C. elegans microRNAs." Mol Cell **11**(5): 1253-1263.
Gregory, R. I., K. P. Yan, et al. (2004). "The Microprocessor complex mediates the genesis of microRNAs." Nature **432**(7014): 235-240.
Griffiths-Jones, S. (2006). "miRBase: the microRNA sequence database." Methods Mol Biol **342**: 129-138.
Griffiths-Jones, S., H. K. Saini, et al. (2008). "miRBase: tools for microRNA genomics." Nucleic Acids Res **36**(Database issue): D154-158.
Grigg, S. P., C. Canales, et al. (2005). "SERRATE coordinates shoot meristem function and leaf axial patterning in Arabidopsis." Nature **437**(7061): 1022-1026.
Grun, D., Y. L. Wang, et al. (2005). "microRNA target predictions across seven Drosophila species and comparison to mammalian targets." PLoS Comput Biol **1**(1): e13.
Gustafson, A. M., E. Allen, et al. (2005). "ASRP: the Arabidopsis Small RNA Project Database." Nucleic Acids Res **33**(Database issue): D637-640.
Hammond, S. M., S. Boettcher, et al. (2001). "Argonaute2, a link between genetic and biochemical analyses of RNAi." Science **293**(5532): 1146-1150.
Han, J., Y. Lee, et al. (2004). "The Drosha-DGCR8 complex in primary microRNA processing." Genes Dev **18**(24): 3016-3027.
Han, J., Y. Lee, et al. (2006). "Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex." Cell **125**(5): 887-901.
Han, M. H., S. Goud, et al. (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.
He, P. A., Z. Nie, et al. (2008). "Identification and characteristics of microRNAs from Bombyx mori." BMC Genomics **9**: 248.
Hiraguri, A., R. Itoh, et al. (2005). "Specific interactions between DICER-like proteins and HYL1/DRB-family dsRNA-binding proteins in Arabidopsis thaliana." Plant Mol Biol **57**(2): 173-188.

Hirsch, J., V. Lefort, et al. (2006). "Characterization of 43 non-protein-coding mRNA genes in Arabidopsis, including the MIR162a-derived transcripts." Plant Physiol **140**(4): 1192-1204.
Hofacker, I. L. (2003). "Vienna RNA secondary structure server." Nucleic Acids Res **31**(13): 3429-3431.
Hofacker, I. L., S. H. Bernhart, et al. (2004). "Alignment of RNA base pairing probability matrices." Bioinformatics **20**(14): 2222-2227.
Hofacker, I. L., W. Fontana, et al. (1994). "Fast Folding and Comparison of Rna Secondary Structures." Monatshefte Fur Chemie **125**(2): 167-188.
Hofacker, I. L. and P. F. Stadler (2006). "Memory efficient folding algorithms for circular RNA secondary structures." Bioinformatics **22**(10): 1172-1176.
Houbaviy, H. B., M. F. Murray, et al. (2003). "Embryonic stem cell-specific MicroRNAs." Dev Cell **5**(2): 351-358.
Hsu, R. J., C. Y. Lin, et al. (2010). "Novel intronic microRNA represses zebrafish myf5 promoter activity through silencing dickkopf-3 gene." Nucleic Acids Res **38**(13): 4384-4393.
Hubbard, S. J., D. V. Grafham, et al. (2005). "Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags." Genome Res **15**(1): 174-183.
Hutvagner, G. (2005). "Small RNA asymmetry in RNAi: function in RISC assembly and gene regulation." FEBS Lett **579**(26): 5850-5857.
Ishizuka, A., M. C. Siomi, et al. (2002). "A Drosophila fragile X protein interacts with components of RNAi and ribosomal proteins." Genes Dev **16**(19): 2497-2508.
Jackson, A. L., S. R. Bartz, et al. (2003). "Expression profiling reveals off-target gene regulation by RNAi." Nat Biotechnol **21**(6): 635-637.
Jacobsen, S. E., M. P. Running, et al. (1999). "Disruption of an RNA helicase/RNAse III gene in Arabidopsis causes unregulated cell division in floral meristems." Development **126**(23): 5231-5243.
Jaillon, O., J. M. Aury, et al. (2007). "The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla." Nature **449**(7161): 463-467.
Johnson, S. M., S. Y. Lin, et al. (2003). "The time of appearance of the C. elegans let-7 microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter." Dev Biol **259**(2): 364-379.
Johnston, R. J. and O. Hobert (2003). "A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans." Nature **426**(6968): 845-849.
Jones-Rhoades, M. W. and D. P. Bartel (2004). "Computational identification of plant microRNAs and their targets, including a stress-induced miRNA." Mol Cell **14**(6): 787-799.
Jones-Rhoades, M. W., D. P. Bartel, et al. (2006). "MicroRNAS and their regulatory roles in plants." Annu Rev Plant Biol **57**: 19-53.
Karolchik, D., R. Baertsch, et al. (2003). "The UCSC Genome Browser Database." Nucleic Acids Res **31**(1): 51-54.
Karolchik, D., R. M. Kuhn, et al. (2008). "The UCSC Genome Browser Database: 2008 update." Nucleic Acids Res **36**(Database issue): D773-779.
Kawamata, T., H. Seitz, et al. (2009). "Structural determinants of miRNAs for RISC loading and slicer-independent unwinding." Nat Struct Mol Biol **16**(9): 953-960.
Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-664.
Kertesz, M., N. Iovino, et al. (2007). "The role of site accessibility in microRNA target recognition." Nat Genet **39**(10): 1278-1284.

Kim, J. H., D. Choi, et al. (2003). "The AtGRF family of putative transcription factors is involved in leaf and cotyledon growth in Arabidopsis." Plant J **36**(1): 94-104.
Kim, V. N. (2005). "MicroRNA biogenesis: coordinated cropping and dicing." Nat Rev Mol Cell Biol **6**(5): 376-385.
Kim, V. N., J. Han, et al. (2009). "Biogenesis of small RNAs in animals." Nat Rev Mol Cell Biol **10**(2): 126-139.
Kurihara, Y., Y. Takashi, et al. (2006). "The interaction between DCL1 and HYL1 is important for efficient and precise processing of pri-miRNA in plant microRNA biogenesis." RNA **12**(2): 206-212.
Kurihara, Y. and Y. Watanabe (2004). "Arabidopsis micro-RNA biogenesis through DICER-like 1 protein functions." Proc Natl Acad Sci U S A **101**(34): 12753-12758.
Lagos-Quintana, M., R. Rauhut, et al. (2001). "Identification of novel genes coding for small expressed RNAs." Science **294**(5543): 853-858.
Lagos-Quintana, M., R. Rauhut, et al. (2003). "New microRNAs from mouse and human." RNA **9**(2): 175-179.
Lagos-Quintana, M., R. Rauhut, et al. (2002). "Identification of tissue-specific microRNAs from mouse." Curr Biol **12**(9): 735-739.
Lai, E. C. (2002). "Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation." Nat Genet **30**(4): 363-364.
Lai, E. C., P. Tomancak, et al. (2003). "Computational identification of Drosophila microRNA genes." Genome Biol **4**(7): R42.
Lanet, E., E. Delannoy, et al. (2009). "Biochemical evidence for translational repression by Arabidopsis microRNAs." Plant Cell **21**(6): 1762-1768.
Lau, N. C., L. P. Lim, et al. (2001). "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans." Science **294**(5543): 858-862.
Lauter, N., A. Kampani, et al. (2005). "microRNA172 down-regulates glossy15 to promote vegetative phase change in maize." Proc Natl Acad Sci U S A **102**(26): 9412-9417.
Lee, M. T. and J. Kim (2008). "Self containment, a property of modular RNA structures, distinguishes microRNAs." PLoS Comput Biol **4**(8): e1000150.
Lee, R. C. and V. Ambros (2001). "An extensive class of small RNAs in Caenorhabditis elegans." Science **294**(5543): 862-864.
Lee, R. C., R. L. Feinbaum, et al. (1993). "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." Cell **75**(5): 843-854.
Lee, Y., C. Ahn, et al. (2003). "The nuclear RNase III Drosha initiates microRNA processing." Nature **425**(6956): 415-419.
Lee, Y., K. Jeon, et al. (2002). "MicroRNA maturation: stepwise processing and subcellular localization." EMBO J **21**(17): 4663-4670.
Lee, Y., M. Kim, et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II." EMBO J **23**(20): 4051-4060.
Lee, Y. S., K. Nakahara, et al. (2004). "Distinct roles for Drosophila DICER-1 and DICER-2 in the siRNA/miRNA silencing pathways." Cell **117**(1): 69-81.
Legeai, F., G. Rizk, et al. (2010). "Bioinformatic prediction, deep sequencing of microRNAs and expression analysis during phenotypic plasticity in the pea aphid, Acyrthosiphon pisum." BMC Genomics **11**: 281.
Leslie, C. S., E. Eskin, et al. (2004). "Mismatch string kernels for discriminative protein classification." Bioinformatics **20**(4): 467-476.
Lewis, B. P., C. B. Burge, et al. (2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." Cell **120**(1): 15-20.

Lewis, B. P., I. H. Shih, et al. (2003). "Prediction of mammalian microRNA targets." Cell **115**(7): 787-798.
Li, R., Y. Li, et al. (2008). "SOAP: short oligonucleotide alignment program." Bioinformatics **24**(5): 713-714.
Li, S. C., P. Tang, et al. (2007). "Intronic microRNA: discovery and biological implications." DNA Cell Biol **26**(4): 195-207.
Li, W. X., Y. Oono, et al. (2008). "The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and posttranscriptionally to promote drought resistance." Plant Cell **20**(8): 2238-2251.
Liang, C., X. Zhang, et al. (2010). "Identification of miRNA from Porphyra yezoensis by high-throughput sequencing and bioinformatics analysis." PLoS One **5**(5): e10698.
Liang, M., E. Davis, et al. (2006). "Involvement of AtLAC15 in lignin synthesis in seeds and in root elongation of Arabidopsis." Planta **224**(5): 1185-1196.
Lim, L. P., M. E. Glasner, et al. (2003). "Vertebrate microRNA genes." Science **299**(5612): 1540.
Lim, L. P., N. C. Lau, et al. (2003). "The microRNAs of Caenorhabditis elegans." Genes Dev **17**(8): 991-1008.
Lin, S. L., J. D. Miller, et al. (2006). "Intronic microRNA (miRNA)." J Biomed Biotechnol **2006**(4): 26818.
Lingel, A., B. Simon, et al. (2003). "Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain." Nature **426**(6965): 465-469.
Lingel, A., B. Simon, et al. (2004). "Nucleic acid 3'-end recognition by the Argonaute2 PAZ domain." Nat Struct Mol Biol **11**(6): 576-577.
Liu, C. G., G. A. Calin, et al. (2008). "MicroRNA expression profiling using microarrays." Nat Protoc **3**(4): 563-578.
Liu, C. G., R. Spizzo, et al. (2008). "Expression profiling of microRNA using oligo DNA arrays." Methods **44**(1): 22-30.
Liu, D., J. Fan, et al. (2009). "Identification of miRNAs in a liver of a human fetus by a modified method." PLoS One **4**(10): e7594.
Llave, C., K. D. Kasschau, et al. (2002). "Endogenous and silencing-associated small RNAs in plants." Plant Cell **14**(7): 1605-1619.
Llave, C., K. D. Kasschau, et al. (2002). "Endogenous and silencing-associated small RNAs in plants." Plant Cell **14**(7): 1605-1619.
Long, D., C. Y. Chan, et al. (2008). "Analysis of microRNA-target interactions by a target structure based hybridization model." Pac Symp Biocomput: 64-74.
Lu, C., B. C. Meyers, et al. (2007). "Construction of small RNA cDNA libraries for deep sequencing." Methods **43**(2): 110-117.
Lu, S., Y. H. Sun, et al. (2005). "Novel and mechanical stress-responsive MicroRNAs in Populus trichocarpa that are absent from Arabidopsis." Plant Cell **17**(8): 2186-2203.
Lund, E., S. Guttinger, et al. (2004). "Nuclear export of microRNA precursors." Science **303**(5654): 95-98.
Ma, J. B., K. Ye, et al. (2004). "Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain." Nature **429**(6989): 318-322.
MacRae, I. J., K. Zhou, et al. (2007). "Structural determinants of RNA recognition and cleavage by DICER." Nat Struct Mol Biol **14**(10): 934-940.
Mallory, A. C., D. P. Bartel, et al. (2005). "MicroRNA-directed regulation of Arabidopsis AUXIN RESPONSE FACTOR17 is essential for proper development and modulates expression of early auxin response genes." Plant Cell **17**(5): 1360-1375.

Mallory, A. C., B. J. Reinhart, et al. (2004). "MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region." EMBO J **23**(16): 3356-3364.
Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.
Mateos, J. L., N. G. Bologna, et al. (2010). "Identification of microRNA processing determinants by random mutagenesis of Arabidopsis MIR172a precursor." Curr Biol **20**(1): 49-54.
McCaig, B. C., R. B. Meagher, et al. (2005). "Gene structure and molecular analysis of the laccase-like multicopper oxidase (LMCO) gene family in Arabidopsis thaliana." Planta **221**(5): 619-636.
Mccaskill, J. S. (1990). "The Equilibrium Partition-Function and Base Pair Binding Probabilities for Rna Secondary Structure." Biopolymers **29**(6-7): 1105-1119.
Meyers, B. C., M. J. Axtell, et al. (2008). "Criteria for annotation of plant MicroRNAs." Plant Cell **20**(12): 3186-3190.
Mi, S., T. Cai, et al. (2008). "Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide." Cell **133**(1): 116-127.
Mica, E., V. Piccolo, et al. (2010). "Correction: High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in Vitis vinifera." BMC Genomics **11**: 109.
Montgomery, T. A., M. D. Howell, et al. (2008). "Specificity of ARGONAUTE7-miR390 interaction and dual functionality in TAS3 trans-acting siRNA formation." Cell **133**(1): 128-141.
Morozova, O. and M. A. Marra (2008). "Applications of next-generation sequencing technologies in functional genomics." Genomics **92**(5): 255-264.
Mourelatos, Z., J. Dostie, et al. (2002). "miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs." Genes Dev **16**(6): 720-728.
Moxon, S., R. Jing, et al. (2008). "Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening." Genome Res **18**(10): 1602-1609.
Moxon, S., F. Schwach, et al. (2008). "A toolkit for analysing large-scale plant small RNA datasets." Bioinformatics **24**(19): 2252-2253.
Nakielny, S. and G. Dreyfuss (1999). "Transport of proteins and RNAs in and out of the nucleus." Cell **99**(7): 677-690.
Ng Kwang Loong, S. and S. K. Mishra (2007). "Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification." RNA **13**(2): 170-187.
Ohler, U., S. Yekta, et al. (2004). "Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification." RNA **10**(9): 1309-1322.
Palatnik, J. F., E. Allen, et al. (2003). "Control of leaf morphogenesis by microRNAs." Nature **425**(6955): 257-263.
Pang, M., A. W. Woodward, et al. (2009). "Genome-wide analysis reveals rapid and dynamic changes in miRNA and siRNA sequence and expression during ovule and fiber development in allotetraploid cotton (Gossypium hirsutum L.)." Genome Biol **10**(11): R122.
Pantaleo, V., G. Szittya, et al. (2010). "Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis." Plant J **62**(6): 960-976.

Park, W., J. Li, et al. (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, A. E., B. J. Reinhart, et al. (2000). "Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA." Nature **408**(6808): 86-89.
Pfeffer, S., A. Sewer, et al. (2005). "Identification of microRNAs of the herpesvirus family." Nat Methods **2**(4): 269-276.
Prigge, M. J. and D. R. Wagner (2001). "The arabidopsis serrate gene encodes a zinc-finger protein required for normal shoot development." Plant Cell **13**(6): 1263-1279.
Provost, P., D. Dishart, et al. (2002). "Ribonuclease activity and RNA binding of recombinant human DICER." EMBO J **21**(21): 5864-5874.
Provost, P., R. A. Silverstein, et al. (2002). "DICER is required for chromosome segregation and gene silencing in fission yeast cells." Proc Natl Acad Sci U S A **99**(26): 16648-16653.
Pusch, O., D. Boden, et al. (2003). "Nucleotide sequence homology requirements of HIV-1-specific short hairpin RNA." Nucleic Acids Res **31**(22): 6444-6449.
Qi, Y., A. M. Denli, et al. (2005). "Biochemical specialization within Arabidopsis RNA silencing pathways." Mol Cell **19**(3): 421-428.
Rajagopalan, R., H. Vaucheret, et al. (2006). "A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana." Genes Dev **20**(24): 3407-3425.
Rand, T. A., S. Petersen, et al. (2005). "Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation." Cell **123**(4): 621-629.
Re, M, G. Pesole, et al. (2009) “Accurate discrimination of conserved coding and non-coding regions through multiple indicators of evolutionary dynamics”. BMC Bioinformatics 10:282
Reinhart, B. J., F. J. Slack, et al. (2000). "The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans." Nature **403**(6772): 901-906.
Rezaian, M. A. and L. R. Krake (1987). "Nucleic acid extraction and virus detection in grapevine." J Virol Methods **17**(3-4): 277-285.
Ribeiro-dos-Santos, A., A. S. Khayat, et al. (2010). "Ultra-deep sequencing reveals the microRNA expression pattern of the human stomach." PLoS One **5**(10): e13205.
Rodriguez, A., S. Griffiths-Jones, et al. (2004). "Identification of mammalian microRNA host genes and transcription units." Genome Res **14**(10A): 1902-1910.
Ruby, J. G., C. H. Jan, et al. (2007). "Intronic microRNA precursors that bypass Drosha processing." Nature **448**(7149): 83-86.
Saito, K., A. Ishizuka, et al. (2005). "Processing of pre-microRNAs by the DICER-1-Loquacious complex in Drosophila cells." PLoS Biol **3**(7): e235.
Saraiya, A. A. and C. C. Wang (2008). "snoRNA, a novel precursor of microRNA in Giardia lamblia." PLoS Pathog **4**(11): e1000224.
Schauer, S. E., S. E. Jacobsen, et al. (2002). "DICER-LIKE1: blind men and elephants in Arabidopsis development." Trends Plant Sci **7**(11): 487-491.
Schloss, J. A. (2008). "How to get genomes at one ten-thousandth the cost." Nature Biotechnology **26**(10): 1113-1115.
Schwab, R. and O. Voinnet (2009). "miRNA processing turned upside down." EMBO J **28**(23): 3633-3634.
Seitz, H. and P. D. Zamore (2006). "Rethinking the microprocessor." Cell **125**(5): 827-829.
Sempere, L. F., N. S. Sokol, et al. (2003). "Temporal regulation of microRNA expression in Drosophila melanogaster mediated by hormonal signals and broad-Complex gene activity." Dev Biol **259**(1): 9-18.

Sewer, A., N. Paul, et al. (2005). "Identification of clustered microRNAs using an ab initio prediction method." BMC Bioinformatics **6**: 267.
Shaw, P. J. and J. W. Brown (2004). "Plant nuclear bodies." Curr Opin Plant Biol **7**(6): 614-620.
Shiohama, A., T. Sasaki, et al. (2003). "Molecular cloning and expression analysis of a novel gene DGCR8 located in the DiGeorge syndrome chromosomal region." Biochem Biophys Res Commun **304**(1): 184-190.
Slack, F. J., M. Basson, et al. (2000). "The lin-41 RBCC gene acts in the C. elegans heterochronic pathway between the let-7 regulatory RNA and the LIN-29 transcription factor." Mol Cell **5**(4): 659-669.
Smith, T. F. and M. S. Waterman (1981). "Identification of common molecular subsequences." J Mol Biol **147**(1): 195-197.
Song, J. J., J. Liu, et al. (2003). "The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes." Nat Struct Biol **10**(12): 1026-1032.
Song, L., M. H. Han, et al. (2007). "Arabidopsis primary microRNA processing proteins HYL1 and DCL1 define a nuclear body distinct from the Cajal body." Proc Natl Acad Sci U S A **104**(13): 5437-5442.
Souer, E., A. B. Rebocho, et al. (2008). "Patterning of inflorescences and flowers by the F-Box protein DOUBLE TOP and the LEAFY homolog ABERRANT LEAF AND FLOWER of petunia." Plant Cell **20**(8): 2033-2048.
Stark, A., J. Brennecke, et al. (2003). "Identification of Drosophila MicroRNA targets." PLoS Biol **1**(3): E60.
Sturm, M., M. Hackenberg, et al. (2010). "TargetSpy: a supervised machine learning approach for microRNA target prediction." BMC Bioinformatics **11**: 292.
Sunkar, R., T. Girke, et al. (2005). "Cloning and characterization of microRNAs from rice." Plant Cell **17**(5): 1397-1411.
Sunkar, R., A. Kapoor, et al. (2006). "Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance." Plant Cell **18**(8): 2051-2065.
Sunkar, R. and J. K. Zhu (2004). "Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis." Plant Cell **16**(8): 2001-2019.
Szarzynska, B., L. Sobkowiak, et al. (2009). "Gene structures and processing of Arabidopsis thaliana HYL1-dependent pri-miRNAs." Nucleic Acids Res **37**(9): 3083-3093.
Tabara, H., E. Yigit, et al. (2002). "The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DExH-box helicase to direct RNAi in C. elegans." Cell **109**(7): 861-871.
Taft, R. J., E. A. Glazov, et al. (2009). "Small RNAs derived from snoRNAs." RNA **15**(7): 1233-1240.
Taft, R. J., C. D. Kaplan, et al. (2009). "Evolution, biogenesis and function of promoter-associated RNAs." Cell Cycle **8**(15): 2332-2338.
Takeda, A., S. Iwasaki, et al. (2008). "The mechanism selecting the guide strand from small RNA duplexes is different among argonaute proteins." Plant Cell Physiol **49**(4): 493-500.
Tam, W. (2001). "Identification and characterization of human BIC, a gene on chromosome 21 that encodes a noncoding RNA." Gene **274**(1-2): 157-167.
Terrier, N., D. Glissant, et al. (2005). "Isogene specific oligo arrays reveal multifaceted changes in gene expression during grape berry (Vitis vinifera L.) development." Planta **222**(5): 832-847.

Vapnik VN. (1998). Statistical Learning Theory Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley: New York
Vaucheret, H. (2006). "Post-transcriptional small RNA pathways in plants: mechanisms and regulations." Genes Dev **20**(7): 759-771.
Vaucheret, H. (2008). "Plant ARGONAUTES." Trends Plant Sci **13**(7): 350-358.
Vazquez, F. (2006). "Arabidopsis endogenous small RNAs: highways and byways." Trends Plant Sci **11**(9): 460-468.
Vazquez, F., S. Legrand, et al. (2010). "The biosynthetic pathways and biological scopes of plant small RNAs." Trends Plant Sci **15**(6): 337-345.
Vermeulen, A., L. Behlen, et al. (2005). "The contributions of dsRNA structure to DICER specificity and efficiency." RNA **11**(5): 674-682.
Wang, H., N. H. Chua, et al. (2006). "Prediction of trans-antisense transcripts in Arabidopsis thaliana." Genome Biol **7**(10): R92.
Wang, X. (2006). "Systematic identification of microRNA functions by combining target prediction and expression profiling." Nucleic Acids Res **34**(5): 1646-1652.
Wang, X., J. Zhang, et al. (2005). "MicroRNA identification based on sequence and structure alignment." Bioinformatics **21**(18): 3610-3614.
Washietl, S., I. L. Hofacker, et al. (2005). "Fast and reliable prediction of noncoding RNAs." Proceedings of the National Academy of Sciences of the United States of America **102**(7): 2454-2459.
Wheelan, S. J., D. M. Church, et al. (2001). "Spidey: a tool for mRNA-to-genomic alignments." Genome Res **11**(11): 1952-1957.
Wightman, B., I. Ha, et al. (1993). "Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans." Cell **75**(5): 855-862.
Wu, F., L. Yu, et al. (2007). "The N-terminal double-stranded RNA binding domains of Arabidopsis HYPONASTIC LEAVES1 are sufficient for pre-microRNA processing." Plant Cell **19**(3): 914-925.
Xie, X., J. Lu, et al. (2005). "Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals." Nature **434**(7031): 338-345.
Xie, Z., E. Allen, et al. (2005). "Expression of Arabidopsis MIRNA genes." Plant Physiol **138**(4): 2145-2154.
Xie, Z., L. K. Johansen, et al. (2004). "Genetic and functional diversification of small RNA pathways in plants." PLoS Biol **2**(5): E104.
Xie, Z., K. D. Kasschau, et al. (2003). "Negative feedback regulation of DICER-Like1 in Arabidopsis by microRNA-guided mRNA degradation." Curr Biol **13**(9): 784-789.
Xie, Z. and X. Qi (2008). "Diverse small RNA-directed silencing pathways in plants." Biochim Biophys Acta **1779**(11): 720-724.
Xu, P., S. Y. Vernooy, et al. (2003). "The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism." Curr Biol **13**(9): 790-795.
Xue, C., F. Li, et al. (2005). "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine." BMC Bioinformatics **6**: 310.
Yan, K. S., S. Yan, et al. (2003). "Structure and conserved RNA binding of the PAZ domain." Nature **426**(6965): 468-474.
Yang, D., F. Buchholz, et al. (2002). "Short RNA duplexes produced by hydrolysis with Escherichia coli RNase III mediate effective RNA interference in mammalian cells." Proc Natl Acad Sci U S A **99**(15): 9942-9947.
Yang, L., Z. Liu, et al. (2006). "SERRATE is a novel nuclear regulator in primary microRNA processing in Arabidopsis." Plant J **47**(6): 841-850.

Yin, J. Q., R. C. Zhao, et al. (2008). "Profiling microRNA expression with microarrays." Trends Biotechnol **26**(2): 70-76.
Yu, B., L. Bi, et al. (2010). "siRNAs compete with miRNAs for methylation by HEN1 in Arabidopsis." Nucleic Acids Res **38**(17): 5844-5850.
Yu, B., Z. Yang, et al. (2005). "Methylation as a crucial step in plant microRNA biogenesis." Science **307**(5711): 932-935.
Zamore, P. D., T. Tuschl, et al. (2000). "RNAi: Double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals." Cell **101**(1): 25-33.
Zeng, Y. and B. R. Cullen (2003). "Sequence requirements for micro RNA processing and function in human cells." RNA **9**(1): 112-123.
Zeng, Y., E. J. Wagner, et al. (2002). "Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells." Mol Cell **9**(6): 1327-1333.
Zeng, Y., R. Yi, et al. (2005). "Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha." Embo Journal **24**(1): 138-148.
Zhang, B. H., X. P. Pan, et al. (2006). "Evidence that miRNAs are different from other RNAs." Cell Mol Life Sci **63**(2): 246-254.
Zhang, H., F. A. Kolb, et al. (2004). "Single processing center models for human DICER and bacterial RNase III." Cell **118**(1): 57-68.
Zhang, X. B., X. F. Song, et al. (2010). "Characteristic comparison between two types of miRNA precursors in metazoan species." Biosystems **100**(2): 144-149.
Zhang, X. H., K. A. Heller, et al. (2003). "Sequence information for the splicing of human pre-mRNA identified by support vector machine classification." Genome Res **13**(12): 2637-2650.
Zhao, C. Z., H. Xia, et al. (2010). "Deep sequencing identifies novel and conserved microRNAs in peanuts (Arachis hypogaea L.)." BMC Plant Biol **10**: 3.
Zhao, J. J., J. Yang, et al. (2009). "Identification of miRNAs associated with tumorigenesis of retinoblastoma by miRNA microarray analysis." Childs Nerv Syst **25**(1): 13-20.
Zhu, J. K. (2008). "Reconstituting plant miRNA biogenesis." Proc Natl Acad Sci U S A **105**(29): 9851-9852.
Zien, A., G. Ratsch, et al. (2000). "Engineering support vector machine kernels that recognize translation initiation sites." Bioinformatics **16**(9): 799-807.
Zuker, M. and P. Stiegler (1981). "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information." Nucleic Acids Res **9**(1): 133-148.

Thanks to:

Prof. Graziano Pesole for giving me the opportunity to learn from him and from the bioinformatics group that he founded.

Carmela Gissi for her availability and for being my Ph.D tutor.

David Horner for what he taught me every day about the wonderful miRNAs world, as Ph.D tutor, and about life, as friend.

Giulio Pavesi for being always what he is: fantastic!

Federico Zambelli, Matteo Chiara and Massimiliano Borsani for grappa, beer, express-pizza and fusion dance…..thanks for being my colleagues, but also, and above all, my friends.

Francesca Griggio and Renato Lupi for sharing biology, bioinformatics, caffè and music with me.

Prof. Gianni Dehò for what he taught me before and during my Ph.D.

My family that, despite the distance, always supported me with LOVE.