Deciphering the diversity of small RNAs in plants: the long and short of it

Frank Schwach, Simon Moxon, Vincent Moulton and Tamas Dalmay

Abstract
RNA silencing is a complex and highly conserved regulatory mechanism that is now known to be involved in such diverse processes as development, pathogen control, genome maintenance and response to environmental changes. Since its recent discovery, RNA silencing has become a fast moving key area of research in plant and animal molecular biology. Research in this field has greatly benefited from recent developments in novel sequencing technologies that allow massive parallel sequencing of small RNA (sRNA) molecules, the key players of all RNA silencing phenomena. As researchers are beginning to decipher the complexity of RNA silencing, novel methodologies have to be developed to make sense of the large amounts of data that are currently being generated. In this review we present an overview of RNA silencing pathways in plants and the current challenges in analysing sRNA data, with a special focus on computational approaches.

Keywords: RNA silencing; bioinformatics; high-throughput sequencing; plant molecular biology

INTRODUCTION
The discovery of RNA silencing phenomena in plants and animals in the early 1990s has dramatically changed our view of RNA and its role in gene expression. RNA silencing is now well established as an important and highly sequence specific mechanism of gene regulation that represses gene-activity at the transcriptional or post-transcriptional level. The highly conserved RNA silencing machinery fulfills a multitude of functions in fungi, plants and animals, ranging from the control of protein-coding gene expression to genome-maintenance and defence against pathogens. Moreover, RNA silencing has received much attention in recent years as a tool for research and therapeutic purposes that makes it possible to manipulate gene expression with a very high degree of sequence specificity in vivo. This specificity is mediated by short non-coding RNAs, known as small RNAs (sRNAs). These 18–30 nt long molecules are, with the exception of the recently discovered Piwi-associated (pi)RNAs in animals, the products of RNAaseIII-type enzymes called Dicers. Substrates for Dicer enzymes, and therefore triggers of RNA silencing, are longer double stranded RNA molecules that may arise due to intra- or inter-molecular base pairing or as a result of an RNA-dependent RNA-polymerase (RdRP) activity. To exert its regulatory function, the double-stranded sRNA must be incorporated into a member of the Argonaute protein family, where one strand is degraded while the other anneals to its specific target by establishing Watson–Crick base pairs. So tethered to the target, the effector complex may
induce target RNA degradation, prevention of mRNA translation into protein or DNA and histone modifications that lead to transcriptional silencing.

Ever since the groundbreaking discovery of sRNAs [1], it has been the goal of many laboratories to decipher the endogenous ‘sRNAome’ of living organisms in order to understand its biogenesis and function. Such efforts have been greatly assisted by recent developments in so-called next-generation sequencing technologies, which are ideal for high-throughput analyses of short sequences [2]. However, even before the onset of high-throughput techniques, it was recognised that plant sRNAs occur in a variety of distinct size classes and that plant genomes give rise to a particularly diverse population of endogenous sRNAs [3–7]. We present here an overview of known classes of plant sRNAs, beginning with the particularly prominent and well-characterised class known as microRNAs (miRNA), and a discussion on the challenges posed to current research by the complexity of plant sRNAomes, with a special focus on computational classification and detection problems.

miRNAs

The first miRNA to be discovered was lin-4 in the nematode worm *Caenorhabditis* elegans [1]. It was initially thought to be a biological peculiarity until a number of additional examples were reported and some of these turned out to be highly conserved from worms to mammals [7–12]. These findings firmly established miRNAs as important regulatory molecules, representing a whole new level of gene regulation that had previously been overlooked. The discovery of the first plant miRNAs in the model species *Arabidopsis thaliana* followed soon afterwards [7]. Since then, the number of reports of new miRNAs from species including multi- and unicellular plants and animals and even viruses has increased rapidly: the central miRNA repository, miRBase, has gone from 218 entries in 2002 to 9539 entries in 2009 [13]. Despite the conservation of the components of the RNA silencing machinery across kingdoms, conservation of individual miRNAs between plants and animals has only ever been reported for a single candidate miRNA family (miR854) [14]. Intriguingly though, these candidates map to transposable elements in both kingdoms and have very unusual precursors, which has led other authors to question their miRNA status [15]. Moreover, even within the clade of *Viridiplantae*, there is conservation of individual miRNAs between mosses and modern land plants but not between single-celled green algae and higher plants [16, 17] which would argue for multiple, independent, origins of miRNAs that all use the same highly conserved RNA silencing machinery in a similar way. In contrast, prokaryotes as well as fungi appear to lack miRNAs altogether.

The double-stranded RNA precursors from which mature miRNAs are excised are the product of intra-molecular base pairing within transcripts that contain imperfect inverted repeat regions. These miRNA precursors are thought to arise from inverted duplications of protein-coding genes [18]. Transcripts of the young miRNA gene would initially form perfectly paired long fold-back structures, giving rise to a multitude of sRNAs with perfect complementarity to the gene of origin. Progressive accumulation of small-scale mutations then leads to sequence divergence in the two arms of the fold-back (hairpin) structure. Modern land plants harbour miRNA genes at all stages of evolutionary progression, indicating that new miRNAs may arise (and be lost) frequently [19]. The more mature examples are often found to be highly conserved between species and have often spawned large families of homologous miRNAs through further duplication events. In contrast, younger miRNAs, that are not yet conserved between many species, are usually encoded by a single gene, and are often found to be expressed at lower levels than the more mature miRNAs. This is partly the reason why non-conserved miRNAs were rarely found before high-throughput sequencing of sRNAs became possible. Applying this novel technology has led to the identification of many novel miRNAs in several species [20–24].

miRNA precursors may initially be several kilobases in length in both animals and plants [25–27] and, in plants, are usually encoded in intergenic regions under the control of their own promoters and transcription factor binding sites [28–31]. The imperfect intra-molecular base pairing of the precursor transcripts appears to give important structural clues for the precise excision of the mature sRNA duplex known as the miRNA/miRNA* pair [32]. In plants, one of the Dicer enzymes (DCL1) is responsible for miRNA precursor pre-processing and excision of the mature miRNA duplex [33], but there is recent evidence for the participation of an additional complex in the precursor processing
HYL1 [34], possibly in cooperation with the zinc finger protein SERRATE [35, 36]. Intriguingly, HYL1 has some similarity to the ‘Microprocessor’, which guides the initial steps of miRNA precursor maturation in animals [34, 37, 38]. In plants, the excised miRNA duplex is methylated at the 3’ end by the methyltransferase HEN1 [39, 40] before being incorporated into an Argonaute protein (AGO1 in Arabidopsis) [41] where one strand, the miRNA*, is degraded. The remaining miRNA strand guides the sequence-specific target interference by annealing to complementary sites on mRNAs [42]. The outcome of miRNA targeting is either translational repression, i.e. blocking of the protein production machinery, or miRNA cleavage at a specific position in the miRNA/target duplex, followed by degradation of the cleaved transcript. The method of regulation is dependent on the level of complementarity between miRNA and target site [43]: a high degree of complementarity leads to mRNA cleavage and degradation, whereas poor complementarity leads to translational repression. The latter is predominant in animals, whereas target degradation has long been thought to be the exclusive mode of action of plant miRNAs. This view may change as evidence emerges for potentially widespread translational repression by miRNAs in plants too [44, 45].

Plant miRNAs have been known for a long time to target transcription factors involved in diverse developmental processes [42]. However, plants also employ miRNAs to regulate gene expression in response to environmental stress, such as pathogen attacks, oxidative stress, dehydration or phosphate and sulphate limitation [46].

HETEROCROMATIN-ASSOCIATED sRNA

Plant sRNA profiles usually contain a small number of highly expressed 21 nt sequences, most of them miRNAs, and a predominant 24 nt size class, made up of a diverse population of sRNAs that are often individually expressed at low levels [6, 47–49]. The 24-nt sRNA-class is also referred to as heterochromatic (hc)RNA due to its frequent association with heterochromatic regions in the genome. In Arabidopsis, hcRNA has been shown to guide transcriptional silencing by DNA- and histone-methylation in a process that employs a number of protein components, including RNA-polymerases, members of the Argonaute and Dicer families as well as DNA and histone methyltransferases [50–52]. The exact hcRNA pathway is complex and there are variations that depend on the nature of the target locus [53–57]. Nevertheless, the following general pathway for hcRNA production and action has been emerging from recent studies: RNA-polymerases POLIV and RDR2 provide substrates for Dicer DCL3, which produces 24-nt sRNAs. A complex containing the Argonaute AGO4 and an RNA-binding protein (KTF1) is guided by the sRNAs to nascent transcripts of another RNA-polymerase, POLV, at the target locus, where it induces the recruitment of DNA methyltransferases and histone modifying enzymes [58, 59].

Once initiated, some loci are able to maintain their methylated status, even if the original trigger is removed, in a process that involves the methyltransferase MET1 [60]. In contrast, a majority of methylated sites appear to require continuous production of hcRNAs, possibly in a self-perpetuating feedback loop, to maintain methylation, as illustrated by a recent high-resolution mapping effort of methylation sites and sRNA alignments in Arabidopsis [49].

Repetitive sequences, including transposable elements and tandem repeats, are among the most prominent targets of RNA-directed DNA methylation, but a number of protein-coding genes and their promoters have also been shown to be affected [4, 3, 6, 47, 61–63]. Transposable elements can compromise genome-integrity by moving into new insertion sites, even to the point of inducing chromosome rearrangements with far-reaching consequences [64]. It is therefore not surprising that these mobile elements are tightly controlled by the silencing machinery [53, 62, 65–71]. Intriguingly, however, recent findings indicate that the silencing constraint against transposable elements may be removed under certain conditions, such as environmental stress [72]. It is tempting to speculate that this behaviour may serve as a last resort to increase genetic diversity by the action of re-activated transposable elements in populations, confronted with a changing environment.

A largely unsolved question, and one that is difficult to address experimentally, is whether hcRNA is exclusively cis-acting. In yeast, an important model for transcriptional silencing phenomena, it was recently reported that a negative regulator of RNA silencing, Eri-1 [73], actively restricts the action of
hcRNAs to their locus of origin and prevents trans-silencing of homologous loci [74]. In plants, artificially introduced exogenous genetic material can certainly direct transcriptional silencing at homologous loci in trans [60, 75–77], but this may not be the case for endogenous hcRNAs. For example, sRNAs with multiple perfect matches to the genome, which should be capable of directing DNA-methylation at multiple sites, were found to be associated with observed cytosine methylation at only some of their potential match sites and are thus likely to induce silencing only at their actual loci of origin [49].

**OTHER ENDOGENOUS sRNA CLASSES**

Another class of endogenous sRNAs that is thought to be cis-acting are the stress-inducible natural-antisense (nat-)siRNAs, which originate from overlapping sense and antisense transcripts [78–80]. Such overlapping opposing transcript pairs are not uncommon in *Arabidopsis* [81, 82] and the gene-products of these pairs often participate in similar processes, which might explain the need for a common regulator [78]. While some nat-siRNAs may also be 24 nt long like the majority of hcRNAs, they differ from the latter in that they cause post-transcriptional target degradation, similar to miRNAs, rather than transcriptional inactivation. In contrast to miRNAs though, nat-siRNAs are cis-acting—they target transcripts from their own genomic region of origin.

Another class of endogenous sRNAs is more similar to miRNAs in their mode of action. The trans-acting (ta-)siRNA guide post-transcriptional degradation at remote targets. The ta-siRNA biogenesis demonstrates another layer of complexity of sRNA interactions: a miRNA-initiates ta-siRNA production by targeting a single-stranded precursor transcript, which in turn becomes a substrate for an RNA-dependent RNA-polymerase [83, 84]. In double-stranded form, the precursor becomes a substrate for a Dicer enzyme (DCL4), whose progressive cleavage in 21 nt intervals leads to the ‘phased’ pattern of mature sRNA production that is a hallmark of ta-siRNA loci [85, 86]. The complexity of the sRNA regulatory network goes even further: ta-siRNAs not only have the ability to target protein-coding genes, they can also trigger another round of ta-siRNA production in remote loci, thus orchestrating an entire cascade of gene regulation [87]. One of the functions of plant ta-siRNAs is the induction of developmental phase changes in diverse and distant lineages including mosses, lycopods and flowering plants [86, 88] and they appear to have been established already in early single-celled green algae [16].

**THE CHALLENGES OF ENDOGENOUS sRNA CHARACTERISATION**

In the early days after the discovery of endogenous sRNAs, efforts to detect and clone these molecules individually from biological samples were time- and labour-intensive. Within a few years though, the scene has changed dramatically, due to the rapid development of next-generation high-throughput sequencing of short sequence tags [2]. As a result, obtaining large-scale sRNA data sets in excess of a million reads has become a routine task and the focus is now on the downstream analysis of this type of data.

One way of classifying sequenced sRNAs is to exploit existing knowledge about the genetic requirements of known sRNA classes. However, obtaining mutant or knock-down plants is not always feasible, particularly when working with non-model organisms. It is therefore important to develop computational methods that may assist in the classification based on the information that is available in any particular experimental system.

Given the frequently observed association of distinct sRNA size classes with distinct silencing pathways, it may be tempting to classify endogenous sRNAs solely by this property. Unfortunately, size alone is not a reliable diagnostic attribute as illustrated by 24 nt long miRNAs [89] and the observation that transposons, under certain conditions, give rise to predominantly 21 nt, instead of 24 nt, sRNAs [49, 90, 91].

The most mature computational sRNA classification strategies today are those aiming to identify miRNAs and their targets. Most published methods use properties of already known miRNAs and their precursors to identify novel members of this class from sequenced sRNA (Figure 1A and 1B), but there are also studies that use genomic sequence data alone to predict miRNA loci ab initio, often in conjunction with cross-species conservation tests [92]. Publicly available software tools have recently appeared, which make miRNA classification and
Figure 1: Visual representation of Arabidopsis sRNA loci and their properties. Genome browser views are generated using high-throughput sequence data from Rajagopalan et al. (GEO accession numbers GSM118372 and GSM118373) [40]. Each sRNA read is represented as a single arrow, the colour of which is dependent on the length of the sRNA sequence (red: 21 nt, green: 22–23 nt, blue: 24 nt). (A) The miR166c locus shows a typical two peak pattern of sRNA alignments corresponding to miRNA (highlighted in red) and lower-abundance miRNA* (highlighted in pink). (B) Secondary structure prediction of the miRNA locus in (A), showing a characteristic hairpin-like miRNA precursor with mature miRNA (red) and miRNA* (pink) sequences highlighted. (C) The ta-siRNA locus TASIA contains predominantly 21 nt sRNAs (track 1: all sRNAs) but the characteristic phasing pattern is difficult to detect by eye. However, a statistical analysis (see main text) reveals a significant P-value ($P = 5.9 \times 10^{-6}$) for a phased pattern (track 2: phased sRNAs only). (D) A hcRNA locus, producing a diverse set of predominantly 24 nt sRNAs. The locus is highly expressed in the flower sample (highlighted in pink) but is almost inactive in the leaf sample (highlighted in yellow), which is consistent with the grouping of these sRNAs into a single functional unit.
target prediction from large-scale data sets more accessible to a wider audience [93–95]. miRCat is a web-based tool that was designed for plant miRNA identification from large-scale sequencing data [95], whereas miRDeep is a package for local installation that was developed to identify animal miRNAs [94], but has also recently been adapted to work with plant data sets [96]. Both programs employ a similar strategy, in that they both map sRNA sequences to putative precursors and score their properties according to observations from known miRNAs. The scoring scheme of miRDeep is based on the expected pattern of miRNA-duplex excision from the stem of the precursor hairpins by Dicer, including the characteristic 2 nt 3'-overhang between miRNA and miRNA* in the miRNA-duplex excised. To increase confidence in miRNA predictions it is important to identify putative miRNA targets in the organism of interest. For this purpose, miRU [93] and a recently added enhanced web-server by the same authors (unpublished, http://bioinfo3.noble.org/psRNATarget) as well as a utility in the UEA sRNA tool kit [95] provide services to identify targets from sRNA data sets by applying rules derived from experimentally verified miRNA–target interactions, which govern the number and distribution of mismatches and the thermodynamic stability of the resulting heteroduplex.

Besides miRNAs, ta-siRNAs are another class of endogenous sRNAs that lends itself well to computational identification, due to the characteristic pattern of mature sRNA production. A popular approach is to test the statistical significance of a ‘phased’ alignment pattern of sRNA sequences at their genomic region of origin (Figure 1C) and to search for the target site of the putative initiator miRNA [87, 95, 97].

Other types of endogenous sRNAs are usually not so well defined by alignment patterns or sRNA sequence properties. It has been known for a long time that Dicer produces sRNAs from long perfect or near-perfect double-stranded precursors in a more random and unpredictable pattern [98]. If it is available, the classification may be based on genome annotation in some of these cases. For example, localisation to overlapping convergent transcript sites is diagnostic of nat-siRNAs [79]. Other endogenous sRNAs are often classified by their co-localisation with repeat elements or gene regions [48]. In fully sequenced model organisms, such as Arabidopsis, it is now also possible to obtain whole-genome methylation maps at high resolution, which may be used to identify hcRNAs that may direct and maintain the observed methylated state at their regions of origin [99, 100].

It is often advisable to combine sRNA alignments by assigning them to genomic loci of origin for classification and comparison purposes. Comparing expression levels of individual sRNA sequences between data sets may be possible for some classes of sRNAs, notably miRNAs, but not for highly complex sRNA loci, where many sRNA sequences are represented by single reads even in high-throughput data sets [48] (Figure 1D). Combining counts of all sequences matching a genomic locus provides a much more robust measurement of sRNA activity at the locus that may be compared between data sets after some transformations of raw counts are applied. First, as in other expression level assays, read counts must be normalised, which is typically achieved by dividing raw counts by the total or genome-matching number of reads in each sample [48, 54]. In addition, a correction for match-repetitiveness has also been suggested, for example by dividing read counts by the number of genomic matches for each sRNA sequence [48, 54]. The rationale behind this transformation is that sRNAs often match highly repetitive genomic sequences, but it is usually not possible to decide which one of its genomic matches is more likely to be the real origin of the sequenced molecule. A recent study has highlighted this problem by demonstrating that pericentromeric repeats actually give rise to relatively low amounts of sRNAs, whereas the opposite result is obtained if no correction for match-repetitiveness is applied [48, 63]. Match repetitiveness is even more problematic when working with unfinished genomes. In these cases, sRNAs or indeed entire sRNA loci may match to sequences that are present more than once in a draft assembly, which may be due to assembly errors rather than true repetitiveness of a genomic region. Vice-versa, the real repetitiveness of an sRNA locus might be underestimated because highly repetitive regions are difficult to assemble and might therefore be under-represented in the draft genome. In addition, even in fully sequenced organisms, it is often not possible to resolve whether sRNAs are actually produced from all of the genomic regions they match to.
One approach for dealing with this particular problem is to identify regions that are true originators of sRNAs by the presence of at least some uniquely-matching sRNAs [48, 54]. Another possibility is to group sRNA loci by sequence similarity and treat each group as one entity for the purpose of the downstream analysis, which has been used to avoid the problems discussed above for unfinished genome assemblies [16].

Another challenge for the identification of genomic sRNA loci is the correct placement of locus boundaries from sRNA alignments alone, which may often be scattered and spurious. One suggested method is based on thresholds of distances between neighbouring sRNA alignments and sRNA abundances [54, 63, 95, 101]. Another approach is to ‘seed’ genomic sRNA loci in fixed-size windows that exhibit significant accumulations of sRNA alignments, followed by merging windows with their neighbours if they contain sRNA alignments [48]. One problem with these methods is the dependence on static thresholds that are often difficult to choose a-priori. However, the expected increase in publicly available high-throughput sRNA sequence data in the near future may be instrumental in improving the detection algorithms and paving the way to a more robust and stable annotation of genomic sRNA-producing loci. This could be done by identifying sRNAs that align to the genome in close proximity and consistently exhibit co-expression patterns across data sets from as many different biological samples as possible, which would indicate that these sRNAs form a functional unit and are most likely derived from a common precursor (Figure 1D).

An in silico analysis of high-throughput sRNA data as described above obviously requires bioinformatics expertise and a suitable computing infrastructure, which may not be available to many laboratories. To solve this problem, publicly available software packages have now started to appear that help researchers to convert sequencer-output into accessible and manageable formats and assist in the data analysis. We have previously published one such package, the UEA plant sRNA toolkit (http://srna-tools.cmp.uea.ac.uk), which is entirely web-based and allows users to process raw large-scale sequencing data and to perform a number of downstream analyses, such as profiling and identification of known and novel miRNAs, filtering of sequences and quantification, grouping and visualisation of sRNA matches to various plant genomes [95].

Another recently published package, CASHX (http://jcclab.science.oregonstate.edu/node/view/56095) [102] has been designed for local installation. The software reads raw sequences from large-scale samples, which are matched to a user-provided genome of interest and the results are stored in a database that can be queried by the user for further downstream analyses.

OUTLOOK

In the last decade, remarkable progress has been made to reveal a previously unknown layer of gene expression regulation involving sRNAs. In plants, this layer turned out to be exceptionally diverse including several different types of sRNAs. The function of these sRNAs is also very diverse, ranging from development to response to environmental changes through genome maintenance and defence against viruses. It is worth mentioning that a similar level of complexity of endogenous sRNAs is beginning to emerge for insects and mammals and many of the approaches currently developed for plant data may be applicable to these systems as well [103–105].

The current challenge is to process these large datasets and decipher biological information from the vast amount of data. Some classes of sRNAs can be identified using existing tools but other classes are still difficult to predict with high confidence. Closely integrated wet-lab and bioinformatics efforts will become increasingly important in developing the methods that are required to decipher the enormous complexity of RNA silencing phenomena in the future.

Key Points

- RNA silencing is a highly complex and conserved regulatory mechanism.
- Novel high-throughput sequencing techniques are instrumental in deciphering the complex sRNA signature in plants.
- While data acquisition is no longer a problem, new methods must be developed to meet the challenges of analysing large-scale data sets of sequenced sRNAs.

Acknowledgements

The authors wish to thank Ericka Havecker for helpful discussions.

FUNDING

Biotechnology and Biological Sciences Research Council (BB/E004091/1).
References


34. Kurth Y, Takashi Y, Watanabe Y. The interaction between DCL1 and HYL1 is important for efficient and precise processing of pri-miRNA in plant microRNA biogenesis. RNA 2006;12:206–12.


41. Baumberger N, Baulcombe DC. Arabidopsis ARGONAUTE1 is an RNA slicer that selectively recruits...


