Computational Methods for High-Throughput Small RNA Analysis in Plants

Monteiro Morgado, Lionel

CHAPTER 1

Introduction

1.1 Small regulatory RNAs: historical milestones

The disclosure of the DNA double helix in 1953 [1] is an important reference for contemporary genetics. It cemented the “one gene-one enzyme” hypothesis formulated in the 1940s [2] and the idea of DNA as a fundamental molecule for life, igniting the hype felt till our times around deciphering the “genetic code”. Although the “one gene-one enzyme” model evolved into a broader concept that encompasses all proteins as “functional” products originated from DNA encoded genes, non-coding regulatory elements have remained widely underestimated. For long time, RNA was mostly seen as a template (messenger RNA) and infrastructural molecule (ribosomal RNA and transfer RNA) matching the archetype of the central dogma of molecular biology defined by Crick in 1958 [3], where genetic information flows from DNA to RNA, and ends translated into a protein.

It was only in the early 1990s that gene silencing was accidentally detected in petunia flowers [4]. Transgenic lines engineered to have a deep purple color as a result of an increased expression of chalcone synthase (a pigment producing gene), resulted in variegated flowers instead. The silencing of the transgene was by then linked to antisense RNA, and because both the transgene and the homologous endogenous gene were affected, the phenomenon was termed co-suppression [4, 5]. A similar mechanism reported in the fungus *Neurospora crassa* was named quelling [6, 7], and a related phenomenon identified in animals was baptized as RNA interference (RNAi). The molecular processes underlying such observations remained unknown for almost a decade. It was only in 1998 that RNAi was described both in plants [8] and *c. elegans* [9]. This last work awarded Andrew Fire and Craig Mello the “Nobel Prize in Physiology or Medicine” in 2006, bolstering the importance of RNAi as an experimental tool and its therapeutic potential. In addition, it was found that post-transcriptional silencing (PTS) can be mediated by another class of RNA molecules with a short length but with a distinct biogenesis: the microRNA (miRNA). The first miRNA was identified in studies with *c. elegans* where *lin-4*, a gene without protein coding capacity, revealed influence in the timing of larval development through the production of tiny regulatory RNA [10]. Only almost a decade later, miRNAs were identified in plants and shown to be involved in various developmental events including flowering [11–13]. In 1994, transcriptional silencing (TS) guided by short length RNA was discovered, introducing the heterochromatic small interfering RNA (hc-siRNA). By then, it was observed in tobacco plants

that viroid RNA can induce methylation of its own complementary DNA, in a mechanism coined RNA-directed DNA methylation (RdDM) [14]. Given the similarities among the regulatory mechanisms described above and the characteristic short size of the RNA molecules involved, they were included under the general designation of small RNA (sRNA). Research with the model organism *Arabidopsis thaliana* added two new elements to the list of sRNAs found in plants. In 2004, the plant-specific trans-acting (ta)-siRNA was described as a secondary product of sRNA activity [15, 16]; and the year after, taught us that naturally occurring antisense transcripts independently synthesized in the genome can hybridize and originate natural antisense transcript (nat)-siRNA [17]. In the meantime, key components for sRNA activity, such as Dicer [18] and Argonaute [19], were identified.

The advent of Next Generation Sequencing (NGS) platforms in the turn of the millennium has permitted to access millions of sRNA sequences in a single experiment [20], showing that some sRNAs can be conserved among species, and that many are tissue and lineage specific. The exact number of sRNAs is unknown but the latest studies suggest that innumerous sequences remain to be characterized [21], many of which may be part of intricate regulatory networks [22] that have been largely overlooked to date.

1.2 Endogenous small RNAs in plants

Although sRNAs can have a source external to the organism where they act, most of the sequences found in natural plants are of internal origin. sRNAs have been shown to have key regulatory functions in development, response to biotic and abiotic stressors, genome stability and transposon control [23]. In plants, sRNA sequences are mostly 21 to 24 nucleotides (nt) in length, and result from cleavage of double-stranded RNA substrates by dicer-like (DCL) enzymes (Figure 1.1). The RNA substrates, themselves, can originate either from a single-stranded RNA precursor with a stem-loop conformation, from overlapping RNA segments produced independently of each other, or from the action of a RNA-dependent RNA polymerase (RDR) over single stranded RNA.

sRNA can be broadly divided into 2 main groups: hairpin-derived sRNA and small interfering RNAs (siRNA). The popular miRNA falls in the first group, while secondary siRNA (including ta-siRNA), nat-siRNA and hc-siRNA fall into the second group. To become active, sRNAs must

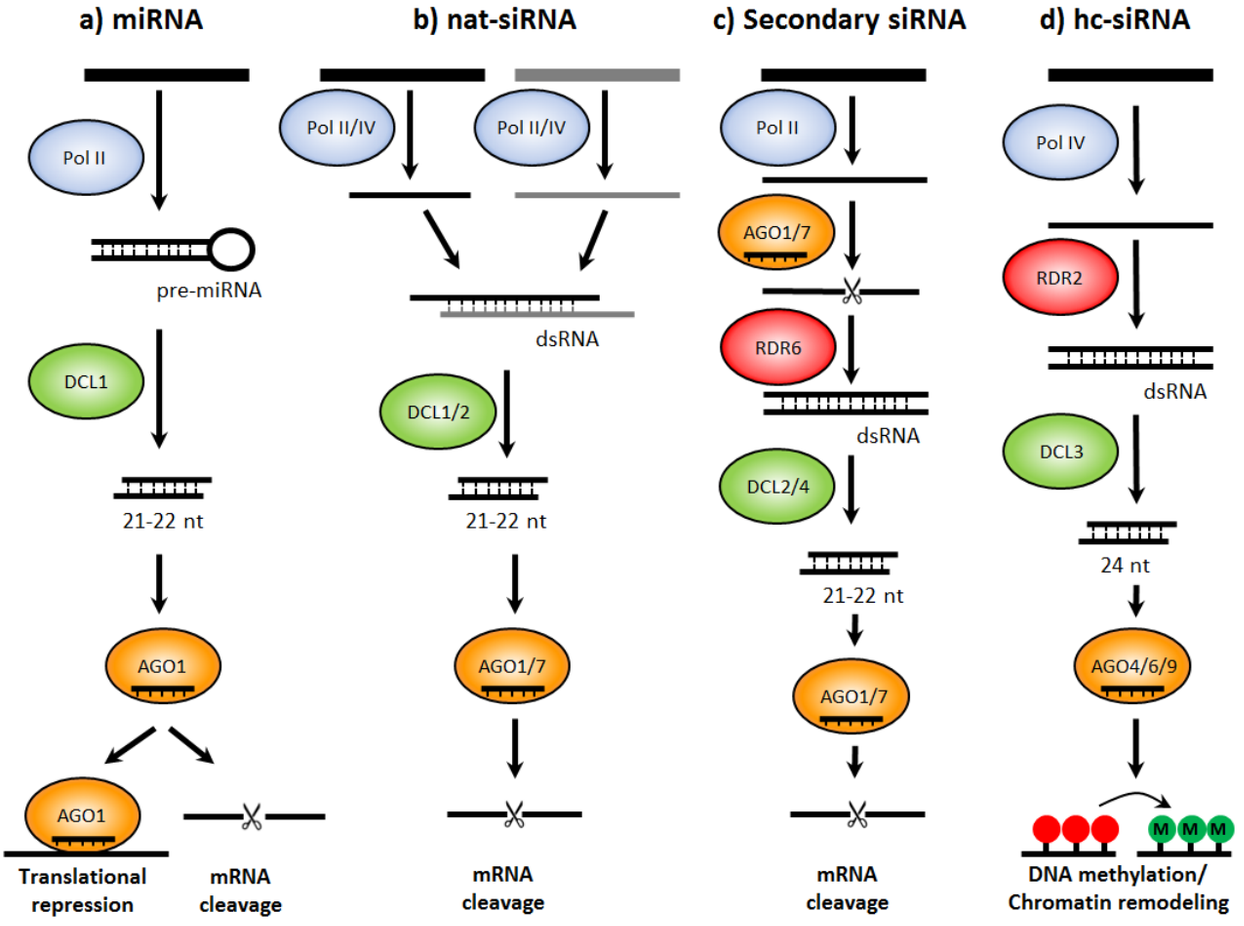

Figure 1.1. Main endogenous sRNA pathways in plants.

load into Argonaute proteins, after which they guide silencing complexes to their targets according to sequence pairing principles. Genomic regulation can then be exerted through TS or PTS. Both modes of action have been intensively studied, but PTS mechanisms, such as messenger RNA (mRNA) cleavage and translation inhibition, are better understood.

1.2.1 Hairpin sRNA and microRNA

A hairpin sRNA (hpsRNA) is defined as a sRNA derived from a precursor that has the capacity to fold into a hairpin-like shape. This precursor can then be processed by any DCL to produce sRNA mostly in the 20-24 nt range. Despite miRNA being an extensively studied subclass under the umbrella of hpsRNA, this parent category is not as well characterized; hpsRNA are more heterogeneous and therefore more difficult to describe [24].

In plants, a primary transcript (pri-miRNA) is produced in the nucleus by RNA polymerase Pol II and further processed into the hairpin-structure precursor (pre-miRNA) by DCL1. A pre-miRNA can be composed of thousands of nucleotides and incorporates inverted repeats that allow the molecule to acquire the hairpin or stem-loop conformation. After folding into a double-stranded structure, DCL1 produces the mature miRNA duplexes. Typically, only one of the strands (the "guide") gives rise to an active miRNA, while the opposite strand (the "passenger") is degraded. Nonetheless, cases where both strands become active have been identified [25, 26].

Although in animals pre-miRNAs typically range from 60 to 80 nt [27], in plants they can be several hundreds of nucleotides in length. Perhaps as a result, the stem-loop structure in plant miRNAs is more variable in size (usually larger) and can contain big bugles. Moreover, in plants the pairing between the arms of the stem-loop conformation shows an overall higher degree of complementarity in the miRNA region [28–30]. Unlike in animals, about 80% of mature plant miRNAs contain a uracil at the 5' ends, which seems to be essential for proper binding and activation of the RNA-induced silencing complex (RISC).

The role of miRNAs in the directed regulation of gene expression is well-known. Individual mature miRNAs, with a characteristic length of 21-22 nt, can target different transcripts but also several miRNAs can target the same transcript [31, 32]. Interestingly, recent evidence indicates that miRNAs can also regulate the concentration of non-coding RNA, including sRNAs [33]. The targeting of non-coding transcripts has been suggested as a mechanism for negative regulation of miRNA concentration in a process called "target mimicry" [34].

miRNAs are thought to be subject to strong evolutionary constraints and tend to be highly conserved across species [13]. Lineage-specific miRNAs differ significantly from more conserved sequences, lacking targets or using unknown non-canonic target criteria, have low abundance, heterogeneous processing from the precursor and are normally encoded by single genes instead of multiple paralogs [24].

1.2.2 Natural antisense transcript siRNA

The main difference between nat-siRNAs and other sRNAs is that nat-siRNAs originate from a RNA duplex formed after the hybridization of a pair of natural antisense transcripts (NATs).

NATs are RNA transcripts that may have or not protein-coding capacity and share complementarity to other RNA transcripts independently produced within the cell [35–37]. Considering the physical distance between NAT producing loci, two main categories emerge: *cis*-NAT and *trans*-NAT. *cis*-NATs are transcribed from the same genomic locus but typically from opposite DNA strands and thus form perfect pairs, while *trans*-NATs are transcribed from distant genomic locations.

Cis-NAT overlapping regions do not have a characteristic length and can occur in five orientations [38]:

1. Head-to-head – consists in the interception in the 5' ends of both transcripts;
2. Tail-to-tail – comprises the interception in the 3' ends of both transcripts;
3. Completely overlapping – a transcript in one strand of the genome is overlapped by the entire length of the other transcript in the opposite strand;
4. Nearby head-to-head – nearby transcripts in a head-to-head manner where the 5'-end of a transcript is near the 5'-end of another transcript in the genome;
5. Nearby tail-to-tail – nearby transcripts in a tail-to-tail manner where the 3'-end of a transcript is near the 3'-end of another transcript in the genome.

Research in multiple species indicates that NATs can assume a variety of significant regulatory roles, such as in alternative splicing [39, 40], RNA editing [41, 42], DNA methylation [43], genomic imprinting [44–49] and animal X-chromosome inactivation [50]; but regulation by NATs is not well understood [51]. In general, *cis*-NATs have very specific targets operating mostly locally and in a one-to-one fashion [52], but *trans*-NATs can share complementarity with multiple transcripts and form complex regulatory networks in processes such as plant response to stress [17]. NAT sequences shared among species such as rice and Arabidopsis have been identified [53], with *cis*-NATs showing high positional conservation between different species [53, 54].

While the molecular requirements for nat-siRNA biogenesis remain elusive, it is clear that the mere existence of a NAT pair is, by itself, insufficient for siRNA production [51]. Accumulation of nat-siRNAs was observed under diverse stressors, positioning this category of sRNA in the adaptation to biotic and abiotic stress [51]. Once annealed, NATs can be fragmented into 21 nt segments by DCL1, enabling PTS; but there are also reports of 24 nt segments produced by DCL3 that may guide DNA methylation [51].

1.2.3 Secondary and trans-acting siRNA

Secondary siRNAs seem to be part of a mechanism to amplify silencing signals intended to affect pathways with a large number of genes, such as in disease resistance and development [55–58]. Two steps describe the core of secondary siRNA biogenesis: the assembly of a double stranded precursor prompted by cleavage of a single stranded RNA segment that afterwards is targeted by RDR to synthesize its complementary strand; and the subsequent processing into siRNA by DCL enzymes. The initial cleavage event is facilitated by a sRNA, after which multiple molecular complexes are recruited for downstream processing. Secondary siRNAs come often from RNA precursors processed by DCL enzymes in a sequential manner from its beginning [59], a phenomenon called phasing. Although not mandatory, this pattern is a strong indicator of siRNA production and has thus been used for secondary siRNA loci detection. Multiple-hit target transcripts are more prone to generate secondary siRNA, especially if susceptible to sRNAs loaded to AGO1 and AGO7 [24].

When targeting involves transcripts produced at distant loci, a secondary siRNA can further be classified as a ta-siRNA. Involvement of ta-siRNA in developmental timing and patterning has been shown [60]. They were first identified in Arabidopsis [15, 16] as a new kind of non-coding RNA that shared similarities with miRNA but that had key differences. Unlike miRNA, a ta-siRNA locus (known as a TAS gene) produces a non-protein-coding transcript that evolves into a double-stranded RNA segment assisted by RDR6. This form is then processed by DCL4 or DCL2, after an initial cleavage event typically facilitated by miRNA, resulting respectively in 21 or 22 nt long RNA segments [61], that can subsequently be incorporated into RISC and direct the cleavage of target mRNA [33, 62].

The ta-siRNA pathway appears to be specific to plants. TAS genes were detected in many plants such as maize [63] and rice [64], and secondary siRNA in general show high conservation among species [59].

1.2.4 Heterochromatic siRNA

hc-siRNAs are typically 24 nt long and mostly derive from transposons, repeats and heterochromatic regions. Their biogenesis is primarily connected to the PolIV-RDR2-DCL3 pathway [65, 66]. hc-siRNA is central for RdDM, which is the pathway responsible for *de novo* DNA methylation. A hallmark of RdDM is the presence of cytosine methylation in all

DNA sequence contexts (CG, CHG, and CHH, where H can be C, A or T) [67, 68]. DNA methylation that is independent of sRNA, by contrast is generally confined to CG and CHG contexts.

There is emerging evidence that the methylation status of plant genomes is altered in response to attack by pathogens [69–71], and that this relies on hc-siRNA production [72–74]. Recent reports connect RdDM with plant immune response through priming of defense genes in processes such as antibacterial resistance [75, 76]. Because DNA methylation can be rapidly reversed by biotic stress, it has been proposed that dampening defense gene expression through active RdDM would provide an effective mode of regulation of host defense responses in plants [77], in a process where defense activation is accompanied by up-regulation of defense genes due to loss of methylation in TEs/repeats in their promoter regions. Such stress-induced changes in sRNA activity can sometimes be transmitted to the progeny in a process termed “transgenerational priming”. This sRNA-mediated process can facilitate quick and transiently adaptive responses, which may be particularly beneficial in fluctuating environments.

1.3 Small RNA in plant transgenerational epigenetic inheritance

The traditional view in genetics assumes that all heritable variation between individuals of a population is encoded in DNA. However, many examples have been reported where DNA alone cannot explain the phenotypic differences observed. Epigenetics, which refers to the study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence [78, 79], has brought new insights into the molecular mechanisms that may underlie this missing heritability. DNA methylation and several histone chemical modifications are widely accepted as epigenetic mechanisms. Some authors further include sRNA in this list as experimental evidence emerges for the transgenerational preservation of meaningful patterns in sRNA populations [80]. However, a stringent definition of epigenetics dictates that those marks must be passed not only through cell divisions but for at least two generations [81]. Such extended criterion aims at discerning epigenetics from what is known as “parental effects”, in which the environment of the parents can influence their offspring. For example, it is known that in some plants the mother has a strong effect on the seeds it produces and that at the time of dispersal, the

offspring is still surrounded by the maternal tissues of the seed coat that play a crucial role for example in the hormonal regulation of seed germination. Because sRNAs can travel long distances inside a plant organism [82] and even be transmitted from the maternal tissues of a plant to the offspring [83], they are considered by many specialists as not being a true epigenetic mark. Either way, their close relationship with other epigenetic footprints like DNA methylation [67], gives sRNAs a solid role as mediators of transgenerational epigenetic inheritance. Compared to animals, in plants the inheritance of epigenetic marks seems more stable. The existence of sRNA-mediated molecular pathways for *de novo* DNA methylation and its maintenance explains such steady transmission. DNA methyltransferases are active during gametogenesis and embryogenesis, allowing plants to evade the reset that happens in animals. Epigenetic inheritance in plants is frequently associated with sRNA originated in transposable elements and that regulates nearby genes [84, 85], which can culminate in inheritable epialleles with long-lasting phenotypic effects. As sessile organisms, plants can benefit by passing to their progeny environmental cues since seeds disperse mostly locally, which increases the chance for new plants to experience environments similar to their ancestors.

As for its relevance for biology and agricultural sciences, transgenerational epigenetic inheritance in plants has attracted much attention, but the vast majority of studies have been focused on DNA methylation, leaving unanswered many questions about the impact of sRNA in such phenomenon.

1.4 Interrelationship between sRNA pathways

If miRNA structure and regulation is quite well studied, the same cannot be said for other sRNA categories. The lack of clear physical features is a significant issue, further aggravated by the fact that many genomic loci can be involved in the biogenesis of sRNA from distinct pathways. It has been noted that some transposon families like Athila, can switch the production of siRNA from 24 nt to 21-22 nt when methylation is lost [86, 87]. This transition starts with the synthesis of transcripts by Pol II that are afterwards degraded into 21-22 nt siRNA. While most of these siRNAs mediate PTS, some can enter a non-canonical RdDM pathway dependent on Pol II, RDR6 and DCL2/4 [87]. This mechanism has the capacity to reestablish TE methylation and correct lost silencing marks that can be further reinforced by

other pathways for DNA methylation. Remarkably, transitions from PTS to TS have also been observed and studied in detail in ÉVADÉ (EVD) [88]. Loss of methylation in this retrotransposon, triggers the production of 21-22 nt siRNA via the action of DCL2 and DCL4. Nevertheless, very high levels of EVD transcripts can saturate the available DCL2 and DCL4, redirecting the siRNA precursors to vacant DCL3 proteins and thus shifting siRNA production from 21-22 nt (mostly) PTS-siRNA to 24 nt hc-siRNA. hc-siRNA can further attenuate the production of PTS-sRNA units via DNA methylation. The numerous cases of epigenetically activated sRNA that have been recently recorded demonstrate clearly the antagonist and complex TS-PTS relationship [22]. It is further known that loss of TE methylation can result in epigenetic activation and consequent transcription, whereby transposon mRNA becomes preferentially targeted by miRNA. To further complicate the equation, the genes for these miRNAs can be controlled by DNA methylation.

It is clear from these examples that plant organisms have large and very intricate sRNA networks connecting genomic loci that may be involved in multiple pathways both as sRNA producers but also as targets. For a clear understanding of sRNA regulation it is therefore fundamental to identify accurately which sequences are involved in which pathways. Only by doing so, we can crack the complex networks formed by these regulatory elements.

1.5 Analyzing sRNA: computational challenges from the "dry lab"

A considerable number of experimental methods for sRNA detection have been developed over the years. Traditional approaches include northern blotting, reverse transcription polymerase chain reaction (RT-PCR), microarrays and more recently NGS protocols like sRNA-seq [89]. Each method has its own limitations, but the use of sRNA-seq has gained momentum in recent years due to its capacity to record in a single run millions of sequences at a genome-wide scale for a relatively low cost. The existence of such large number of sequences makes impractical the application of other experimental assays to determine additional properties for each of the sequences captured under the current technology. For example, gene-specific experimental validation of PTS targets using well established methods like real-time reverse transcription polymerase chain reaction (qRT-PCR), luciferase reporter assays and western blot, are labor intensive and not easily scalable to genome-wide studies [90, 91]. In the case of TS, direct experimental target validation has only been

proposed recently [92] and therefore it is still in its infancy. This is where computational methods come into play, aiding the extraction of additional information from sRNA primary structure, and prioritizing candidates for experimental validation. Nevertheless, the computational analysis of sRNA-seq data is far from being a trivial task. Libraries are composed of mixtures of sRNA sequences with diversified origin, structural properties and involved in distinct modes of silencing. In addition, the identification of a sRNA sequence by itself is not a guarantee of sRNA regulation since cells can produce inactive short-length RNA. Despite the great interest in separating those units which can guide silencing from other sequences, no computational methods for general high-throughput function detection existed prior to the work reported in this thesis. Using the existing software tools, function detection could only be inferred partially for PTS target prediction, implying that a large share of sRNAs which often guides TS remained ignored.

Once functional sRNAs are identified, determining the mode of silencing for each sequence is the next step. Doing so, gives important clues about the duration/transmission of silent states and the type of biological pathways that induce/maintain these states. PTS is interpreted as a faster response than TS, but with a less lasting effect. Again, no computational methods were publically available to determine sRNA-mediated TS by the time the work of this thesis started. In addition, it is well known that tools for PTS target prediction are characterized by high false positive rates, which complicates laboratory validation. Software to distinguish TS- from PTS-sRNA can in principle improve PTS inference. Although the central role of plant Argonaute (AGO) proteins in defining the mode of silencing is widely accepted and frequently mentioned in sRNA research, no computational methods have been devised to explore the potential of a given sRNA sequence to associate with a specific AGO. Determining such capacity has been strictly dependent on sRNA sequencing after AGO immunoprecipitation.

Another limitation for sRNA studies is the high variability found within and across sRNA species, as well their dynamic nature which impacts our ability to capture them in the lab. Indeed, the detection of sRNA can be experimentally difficult as their expression can be low or dependent on specific developmental stages, cell types or stimulation. *In silico* analysis of artificial sRNA (artsRNA) can give preliminary answers to theoretical models and motivate experimental validation. The exploration of artsRNA is not a new concept, as conservation principles have been used in the past to detect miRNA [93]. This philosophy can be revised

to develop programs for general sRNA detection in an era where new genomes are published on a daily basis.

As seen so far, mining sRNA-seq data involves innumerous analytical steps and can even go beyond the inspection of the mature sRNA sequences, precursors and targets. Studies regarding variation in sRNA abundance among samples and many other downstream analyses (e.g.: gene ontology, network analysis, etc.) are common practice. Currently, there are many individual programs for very specific sRNA-related applications dispersed over the internet, but only a small number of integrative frameworks exist. These often emphasize miRNAs and ignore other important sRNA categories. This is especially evident in plants where only a handful of (frankly incomplete) software platforms can be found. Biologists are often not acquainted with programming and lack the necessary computational skills to build the pipelines they need, which culminates in a tremendous time investment to achieve only modest solutions. There is a clear and urgent demand for novel computational tools to empower sRNA researchers. This PhD thesis has aimed to meet this demand by delivering novel computational methods to improve sRNA detection and characterization. These novel tools are embedded in a single computational platform and integrated with existing software programs, thus facilitating comprehensive and flexible analyses of sRNA in plants.