Deep sequencing of small RNAs in plants: applied bioinformatics

David J. Studholme

Advance Access publication date 19 December 2011

Abstract
Small RNAs, including microRNA and short-interfering RNAs, play important roles in plants. In recent years, developments in sequencing technology have enabled the large-scale discovery of sRNAs in various cells, tissues and developmental stages and in response to various stresses. This review describes the bioinformatics challenges to analysing these large datasets of short-RNA sequences and some of the solutions to those challenges.

Keywords: RNA; silencing; microRNA; miRNA; next-generation sequencing; small interfering RNA; siRNA

WHAT IS SMALL RNA?
The discovery of silencing RNAs in the early 1990s, along with several other lines of evidence, has contributed to the need to substantially refine the central dogma of molecular biology (‘DNA makes RNA makes protein’) [1]. It is now clear that RNA plays a variety of centrally important roles apart from being an intermediate in sequence information from DNA to protein. The transcriptome of a plant cell contains many RNA molecules, apart from well characterized players, such as tRNA and rRNA, which are not destined for translation into protein. These range from short molecules such as miRNAs (around 21 nt) to long non-coding transcripts that can be many thousands of nucleotides in length [reviewed in 2, 3]. In this review, we focus exclusively on small RNAs; that is RNA molecules shorter than about 30 nt. We pay particular attention to bioinformatics considerations of analysing large sequence-based datasets derived from these sRNAs. A major aim of this manuscript is to offer newcomers to the field some pointers for solving the bioinformatics challenges associated with large-scale sequencing projects. Our aim is not to extensively review the biology of sRNA in plants nor the history of their discovery; that has been done extensively elsewhere [2–27].

Probably the best understood class of sRNA is the miRNAs. They are typically between 20 and 24 nt in length with most being 21 or 22. A subset of miRNAs tends to be evolutionarily conserved among related organisms and are important post-transcriptional regulators of gene expression [4]. However, the extensive differences in miRNA repertoires between even closely related species suggest that new families of miRNA are frequently spawned and subsequently lost [25]. The biologically active mature miRNA molecule is derived from a longer imperfectly complementary primary miRNA transcript (pri-miRNA), which is transcribed from the genomic DNA by RNA polymerase II [28, 29] or, at least for some human miRNAs, by RNA polymerase III [30]. The pri-miRNA has a hairpin structure that may be up to several kilobases long [26–28]. This pri-miRNA undergoes two cleavage events. In animals, proteins Drosha and DGCR8 (also known as Pasha in flies) catalyse the first cleavage in the nucleus, which results in removal of the 3′ poly-A tail and the 5′-terminal cap to generate the pre-miRNA, which has the characteristic stem–loop structure [31–35]. Exportin 5 mediates the export of the pre-miRNA out of the nucleus and into the cytoplasm [36] where the Dicer RNase III enzyme makes pairs of cuts and releases the short-RNA...
This duplex consists of a pair of strands of 20–24 nt in length (most commonly 21 nt) and has 2 nt 3’ overhangs [38, 39]. One of the two strands comprises the mature active miRNA. The other, complimentary strand is known as miRNA* (or miRNA-star). Finally, the miRNA* is preferentially degraded and the single-stranded mature miRNA is retained. In plants, miRNAs probably evolved independently of animal miRNAs [25] and their biogenesis exhibits some differences from the biogenesis of animal miRNAs [41]. Specifically, in plants, there is no orthologue of DGCR8/Drosha; instead, the Dicer-like protein DCL1, in conjunction with HEN1 and HYL1, orchestrates both cleavage steps (pri-miRNA to pre-miRNA and pre-miRNA to duplex) [42–47]. This biogenesis of miRNA from a fold-back structure via an overhanging duplex has implications for the computational detection of miRNAs in sequence data, as we shall see below.

The miRNAs are just one of several distinct classes of sRNA [5]. In plants, but not in animals, the most abundant class of sRNA are the heterochromatic (hc)RNAs, so called because of their association with heterochromatic regions of the plant genome. These are most commonly 24 nt in length and show great sequence complexity, with many thousands of different RNA sequences being present in a single type of cell or tissue [48–51]. Each individual species is expressed at a low level but collectively the 24-nt hcRNAs may account for approximately half of a plant’s total mass of sRNA. Although they resemble miRNAs, being single-stranded RNA molecules with overlapping distributions of length, siRNAs originate from a different biogenesis pathway and have very different functions [52–59]. It has been shown that hcRNAs play a role in methylation of DNA and histones at specific regions of the genome leading to silencing of transcription. The RNA polymerases POLIV and RDR2 transcribe single-stranded hcRNA precursors, which are cropped to 24 nt by Dicer-like ribonuclease 3 (DCL3) [60, 61].

Several other classes of sRNA have been discovered recently that have important biological functions, though they are generally much less abundant than miRNA and hcRNA. These include trans-acting short-interfering RNAs (tasiRNA) [62, 63] and natural-antisense transcript short-interfering RNA (NATsiRNA) [64–66]. There are likely other classes yet to be discovered. The term small interfering RNA (siRNA) is often applied to any sRNA with a length in the region of 19–25 nt that is generated by any of the Dicer-like ribonucleases DCL2, DCL3 and DCL4, but excluding the miRNAs, which are generated by DCL1 [3, 20, 21, 35].

In recent years, miRNAs and the various classes of siRNA have emerged as important players in plant development and responses to abiotic and biotic stress [7, 8]. They probably also play a role in systemic signalling within the plant [9, 67]. This has motivated numerous studies aimed at discovering and cataloguing the populations of sRNA present in various tissues and developmental stages and under treatments both in the model plant Arabidopsis thaliana and in trees, crops and other plants [48, 50, 51, 68–80] as a step towards understanding their function. It has been argued that much of this research is purely descriptive and does not provide mechanistic insights [41]. However, deep-sequencing of sRNAs in plants has been a key tool leading to several important discoveries. For example, large-scale sequencing of sRNAs in Arabidopsis revealed functional interconnections between sRNA and DNA methylation [51]. Comparison of sRNA repertoires between wild-type and mutants revealed a correlation between RDR2- and DCL3-dependent 24-nt sRNAs and repetitive elements in the Arabidopsis genome [48], providing a strong clue to the functions of this class of sRNAs.

There are broadly two approaches to sRNA discovery. The first is ab initio identification or prediction of expressed sRNA sequences in the genome sequence. The second is a more empirical approach of determining the sequences of sRNA molecules isolated from biological material [81, 82].

**SOME sRNA CAN BE PREDICTED FROM GENOME SEQUENCE**

Until the advent of massively parallel sequence signature sequencing (MPSS) [81] and subsequent next-generation DNA sequencing technologies, the only options for discovery of previously unknown sRNA species were genetic screening [83, 84], Sanger capillary sequencing of sRNA libraries [85–87] or bioinformatic analysis of available genome sequence data [82, 88–92]. Most of these latter bioinformatics approaches focused on miRNAs [reviewed in 82, 88, 89]. The simplest
approach is to identify homologues of known miRNAs in the same genome or in other species using nucleotide sequence similarity searches [47, 91, 92]. It is possible to identify novel miRNAs with no known homologues by exploiting the following features: (a) some miRNA sequences are highly conserved among related species [93] and (b) miRNAs are derived from a precursor that has a hairpin structure [93]. The expected length of the precursor sequence in plants is typically between 80 and 320 nt [94]. The essence of the approach is to scan intergenic sequences of sets of two or more related genomes for highly conserved sequences falling within an imperfect inverted repeat. At the time that these predictive studies were performed, around 2003–2005, the only complete plant genome sequences available were for A. thaliana and rice [94–97]. Subsequent bioinformatics studies used genomes of other plants as they became available. Where complete genome sequence was not available, some researchers searched the transcriptome (expressed sequence tags, ESTs) and/or incomplete genomic sequence data [98–105]. It should be noted that this approach may be less successful when applied to plants than to animals because of greater variability in length (typically 60–400 nt) and in secondary structure in plant miRNA precursors [5, 97].

A third type of approach has also been developed that does not rely on genome comparisons nor on sequence similarity searches with known miRNAs [106, 107]. This approach exploits the near complementarity between the sequences of mature miRNAs and their regulatory targets. It considers all genomic segments of around 20 nucleotides that are complementary to a target transcript (up to two mismatches). These sequence segments are subjected to a series of filters to remove low-complexity sequences, unstable miRNA precursor duplexes and long sequence matches [106, 107].

### SMALL RNA CAN BE DETECTED EMPIRICALLY THROUGH SEQUENCING

The bioinformatic methods for predicting miRNA in genome sequences have been quite successful and have identified numerous candidates that have subsequently been confirmed experimentally [93, 95–97]. However, the predictive method is limited by high rates of false positives and false negatives, which can only be eliminated by laborious and expensive follow-up experiments [41]. Furthermore, these predictive methods are obviously not very helpful for identifying siRNA and other classes of RNA that are not derived from hairpin-structured precursors. Neither are sequence similarity-based approaches useful for identifying the large numbers of recently evolved miRNAs that are not conserved among related species [25]. Therefore, there is a role for more empirical approaches to discovering sRNA. In the days before cheap high-throughput DNA sequencing [108], this meant cloning short cDNA (derived from sRNA) and sequencing by Sanger capillary sequencing [85–87]. This was, of course, relatively laborious and expensive. Furthermore, it was strongly biased towards the most abundant sRNA species, which mostly comprised a small subset of miRNAs [25, 109]. However, in 2005 the then new technology of massively parallel sequence signatures (MPSS) was used to generate more than 2 million sRNA sequence signatures from A. thaliana flowers and seedlings [110]. This was a major leap forward, the depth of the data providing the first glimpse of the sheer diversity of sRNA molecules present in plants; there were more than 1 million distinct sequence tags present. The few sRNA sequences that had previously been described accounted for just the tip of the iceberg, with the discovery of more than 90 000 distinct sRNA sequence signatures [110].

Ground-breaking as they were, the MPSS-based studies of plant sRNA were limited by the fact that the resulting sequence signatures were of a short and fixed length (typically 17 or 20 nt), leading to some ambiguity about the full-length sequences of sRNAs [110]. By 2006, the first of the so-called next-generation sequencing technologies had become commercially available. This was the 454 GS10 pyrosequencing platform, which would generate around 100 000 sequence reads of up to 100 nt long [111]. This read length was more than sufficient for sequencing the entire lengths of sRNA molecules. Several laboratories quickly exploited this high-throughput sequencing platform to survey sRNAs in the model plant A. thaliana [78 112–117], in other plants [69, 118] and in the single-celled green alga Chlamydomonas [119, 120]. Note that RNA was not sequenced directly; it was first ligated 3’ and 5’ adapters and reverse-transcribed and amplified. Then the cDNA, including the
transcribed adapter, was subjected to pyrosequencing [48, 119]. In the resulting sequence reads, the 5’-end of the read corresponded to the 5’-end of the sRNA and the 3’ adaptor sequence marked the 3’-end of the sRNA [121]. So now we had a high-throughput platform, which did not rely on cloning, that could elucidate the complete sequences of hundreds of thousands of sRNAs. By incorporating ‘bar-code’ sequences into the 3′ adapters, we could even distinguish between different sequencing libraries sequenced simultaneously in the same pyrosequencing run (i.e. multiplexing) [122]. For example, Mosher and colleagues [78] used barcoded cloning and adapter primers to allow the combination of sRNA libraries from wild-type and two mutants within a single pyrosequencing run.

The next major step forward came in 2007 when a second major next-generation sequencing technology became available. Illumina’s Solexa GA technology [123] could generate sequences for more than half a million sRNA molecules per lane (where one run consisted of a flowcell containing eight lanes). Compared with the 454 platform, this represented at least a 100-fold reduction in cost per nucleotide, quickly making this the platform of choice for sRNA discovery [124–131]. The sequence reads from Illumina’s GA were shorter than those of the 454 GS20 (36 nt versus 100 nt) but this was not a problem for sRNAs. Subsequently the Illumina GA2 has been superseded by the GA2 and the HiSeq 2000 with ever increasing rates of data generation and falling costs per nucleotide. Currently, one can expect many tens of millions of reads per lane. Recently, the ABI SOLiD platform [132] has also come onto the market and offers comparable read lengths and per nucleotide costs to the Illumina. The first SOLiD-based sRNA studies were published in 2010 and 2011 [133–149] including one study of the plant Arabidopsis lyrata [149]. A further recent development has been the launch of several relatively cheap and physically smaller high-throughput sequencing instruments including Life Technologies’ Ion Torrent [150], Illumina’s MiSeq and Roche’s GS Junior [151]. Although these platforms offer significantly lower throughput and significantly higher per-base monetary cost, their low initial cost of ownership is likely to gain them a foothold in a wide range of research laboratories, and their relative portability may eventually earn them a place in diagnostics laboratories in such settings as agricultural or military field stations.

### MAKING SENSE OF sRNA THROUGH MAPPING TO A REFERENCE GENOME

The first steps in dealing with sRNA sequencing datasets are usually to remove any adapter sequences and obtain an overview of data quality. For these tasks, the FASTX Toolkit is invaluable [http://hannonlab.cshl.edu/fastx_toolkit/links.html]. This toolkit includes a range of utilities for trimming and filtering sequences reads in FastA or FastQ formats generated by 454 and Illumina high-throughput sequencing platforms. It also offers tools for generating graphical overviews of the distributions of sequence quality scores. When manipulating sequence data with a range of different bioinformatics software, it is often necessary to convert between file formats. For example, there are at least three different flavours of the popular FastQ file format [152], which is commonly encountered as the output data from Illumina sequencing platforms. The FASTX Toolkit facilitates conversion between the various FastQ formats and also between FastQ and the very widely used FastA format.

A substantial proportion of the sRNA library is often comprised of turn-over products of larger RNAs such as tRNA and rRNA, so many studies [50, 70, 78, 79, 119] have filtered these on the basis of sequence similarity to a database of such sequences [153–155]. Once these preliminaries have been dealt with the next step is usually alignment of the sRNA sequences against a reference genome sequence.

One good reason for aligning reads against a reference sequence is to eliminate erroneous data. No sequencing platform is error-free [151, 156, 157]. Assuming an error-rate of 1% per nucleotide, and an even distribution of errors along the length of the read, then for a population of 21-nt sequence reads, we would expect at best only about 81% of sequences to be free of errors. Of course, these assumptions are unlikely to be correct, but the important point is that some proportion of the sequence reads will inevitably contain errors. In practice, especially in some of the earlier datasets, error-rates could be much higher [158–160].

Sequence reads containing one or more errors are dangerous because they could be misinterpreted as newly discovered species of sRNA. A straightforward method for eliminating these error-containing reads is to align against the relevant reference genome sequence and discard all sequence reads
that do not match with 100% identity [31, 68, 69, 73, 78, 79, 119] or that have more than one mismatch [80]. Other studies have attempted to ‘repair’ reads containing a single mismatch where this could be achieved unambiguously by substituting a single nucleotide [48]. In the case of studies involving plant responses to challenge by a pathogen, then the pathogen genome should be included as a reference [80]. Of course, this approach is not without pitfalls. In theory, an erroneous read might, by chance, misleadingly align to an alternative near-identical sequence in the genome. Conversely, a perfectly accurate sequence read may fail to align because of errors in the reference genome sequence, resulting in wasting of some of the data. Similarly, if the individual plant from which the sRNAs are collected is genetically different from the best available reference sequence, then perfectly accurate sequences may be discarded unnecessarily. Insisting on 100% identity to the genome also assumes that sRNA is transcribed from the genome with perfect fidelity and absence of any post-transcriptional modifications, which are known to occur in at least a sub-population of sRNA molecules [161, 162].

But what if no genome sequence is available? ESTs or genomic survey sequences may be available for use as a reference [163]. However, such reference sequences are likely to be of lower accuracy than a finished genome sequence and will likely contain errors and redundancy.

Several high-quality software packages are freely available for aligning large sets of short nucleotide sequences against a reference genome sequence (reviewed in [164]). Irrespective of the choice of alignment tool, the user is faced with the problem of cross-mapping, which is the situation in which an sRNA sequence originating from one genomic location is inadvertently mapped to another, incorrect, location [165]. Cross-mapping arises as the result of multiple occurrences of identical sequences within the genome. Most sRNA sequencing studies have dealt with the problem of cross-mapping in an essentially arbitrary fashion, for example, randomly assigning sRNAs to the various match sites. However, a method has recently been proposed that uses the density of nearby genomic sites matching sRNAs in order to infer the most likely ‘true’ site of origin for each cross-mapping sRNA [165]. The problem of cross-mapping inevitably arises with sRNAs that originate from inverted repeats in the genome sequence as is the case for many h-cRNAs. In one major pyrosequencing survey of sRNA transcribed by RNA polymerase IV [113], this problem was tackled by considering only those reads that overlapped imperfectly repeated sites and were thus able to be distinguished.

As an alternative to filtering sequence reads on the basis of identity to a reference sequence, one could filter reads on the basis of their base-call quality scores [166]. However, even base calls with the highest quality scores are not guaranteed to be free of errors [158]. Schreiber and colleagues devised a novel approach to overcome sequence errors [72]. They used Illumina sequencing to survey miRNAs in barley, an important crop plant for which no complete genome sequence is yet available. They proposed a method for identifying error-containing sequence reads by identifying low-abundance sequence variants that were nearly identical to abundant sequences. They propose a model whereby sequencing errors occur randomly with a given probability. Then it is straightforward to calculate the expected frequency for sequence variants containing one or more errors. They found that their data fitted their model very well, supporting their proposal to identify erroneous reads on the basis of relative abundances of clusters of similar sequences.

There are other good reasons for aligning sequence sRNA data against a reference genome, aside from eliminating sequencing errors. The majority of the siRNAs are expressed at low levels, meaning that they cannot be quantitatively compared between different biological samples; the numbers of observed sequence reads are so low that differences among biological samples are within the range of variability between biological replicates [167]. However, on aligning against the genome, it becomes apparent that the sRNAs are not distributed uniformly nor randomly, but rather tend to be much more aggregated than would be expected by chance. Assuming that sRNA sequences are faithfully transcribed from the genomic DNA, aligning the sRNA against the genome can reveal the genomic locus or ‘gene’ that encodes that sRNA [168, 169]. In the case of miRNAs, the structure of the genetic locus is relatively simple. It encodes a hairpin-structured transcript, within which a single subsequence often dominates and represents the mature miRNA. Many or most of the observed sRNA sequence reads originating from this locus will be identical to each other [3], though there may also be detectable quantities of the complimentary miRNA*.
also be significant numbers of sequence reads that correspond to isomiRs [170], which are mature variants of the miRNA [171]. However, the sequences of the isomiRs overlap most of their lengths [170, 171]. This low sequence complexity means that it is relatively easy to identify differences in miRNA profiles between different biological samples. In contrast, an siRNA-generative locus may be associated with tens or hundreds of different siRNA molecules, each of which is not very abundantly expressed. In this case, it is only possible to make sense of the data by clustering sets of sRNA together according to their genomic loci of origin [168, 169].

The problem that we need to solve is how to identify and delineate biologically meaningful genomics segments or loci. Then, rather than comparing abundances of individual siRNAs between biological samples, we should compare the total abundances of siRNA from the genomic locus.

It is not obvious how best to delineate the boundaries of sRNA-generating loci in a biologically meaningful way and a common approach has been to simply divide the genome into a series of equally sized ‘clusters’ or ‘bins’ [172]. Another common, and equally arbitrary, approach to defining sRNA-generative loci is based on identifying maximally contiguous regions of the genome with an average alignment density above some defined threshold. This is the approach that we used in [78, 119] and is implemented in the SiLoCo tool, freely available as part of the UEA sRNA Toolkit [173]. When calculating the alignment density, each aligned sRNA sequence has its weight adjusted according to the degree of confidence that it has been aligned to the correct position on the genome; reads that share perfect sequence identity with multiple genomic sites will be given less weight than those that align uniquely to a single site. This approach has the advantages of being relatively simple to understand and to implement. A major disadvantage is that the outcome is very sensitive to the choices of parameter values (for example, the minimum depth threshold and the weighting scheme) and to the quantity of input data.

In an attempt to overcome the limitations of the SiLoCo-based method, we developed a graph-based algorithm for inferring sRNA-generative loci from alignments of sRNA sequences against a reference genome. This method gave excellent robustness to varying values of its two parameters and sensitivity [169]. Other promising approaches include time-series data-mining algorithms [174] and Bayesian methods [168]. Currently, it remains unclear which of these approaches most closely approximates biological reality. It is likely to remain unclear until we have a fuller understanding of all the mechanisms that give rise to siRNA and we have an extensive catalogue of siRNA-precursor transcripts under a broad range of biological scenarios.

DIFFERENT CLASSES OF sRNA GIVE DISTINCT PATTERNS OF SEQUENCE ALIGNMENT

Genomic loci associated with different classes of sRNA show very different patterns of aligned reads; see Figure 1 in [3]. For example, miRNA loci usually show a distinct narrow peak corresponding to the mature miRNA and its isomiRs [170, 171], often a second peak corresponding to the miRNA* and very little background signal. Loci-encoding tasiRNA show a very different pattern. At a tasiRNA locus, 21-nt sRNAs align in a phased arrangement whereby the majority of sequences initiate at a series of evenly spaced sites [3, 175]. The non-random spatial distribution of the aligned sequences means that it is relatively easy to detect tasiRNA loci in a genomic alignment using a statistical approach; specifically, the method involves searching for segments of the alignment in which the distribution of the start sites significantly deviates from a hyper-geometric distribution [175]. The UEA sRNA Toolkit includes a slight modification of the original [175] method in that it takes into account lengths of the input sequences and sets a lower limit on the abundance of sequences that it considers.

Both miRNAs and ta-siRNAs are associated with very distinctive patterns of aligned sRNAs. The majority of sRNA-generative loci do not match these patterns and give rise to complex mixtures of sRNA species that appear to be randomly selected sub-sequences from the locus [3]. However, on close inspection, there do appear to be patterns. For example, many sRNA loci show enrichment for a particular size-class of sRNA and/or a bias towards sRNAs originating on one strand more than the other. It is possible that these differences in empirical properties might reflect underlying biologically significant differences, such as different biogenesis pathways. Therefore, there might be great value in clustering or classifying loci according to a panel of
empirical properties in order to discover new meaningful sub-classes of siRNA. Of course, the major challenge is in the choice of properties (i.e. ‘feature selection’) and the choice of weighting accorded to each property.

**TO MANAGE DATA, USE A DATABASE MANAGEMENT SYSTEM**

Data management is an important issue for any laboratory that generates more than a few sRNA high-throughput sequencing datasets. It is quite possible to subject each dataset to a standard workflow. This workflow might include steps such as adapter removal, quality filtering, alignment against a reference genome, locus detection, quantitative expression profiling, miRNA discovery, etc. [3, 173, 176, 177–179]. However, to maximize the value of the data, it is useful to cross-reference against previous data from the same and/or other laboratories. While providing bioinformatics support for various large-scale sRNA sequencing projects at The Sainsbury Laboratory [78, 79, 119], we developed an automated sRNA analysis system that integrated the standard workflow with a comprehensive database of previous data (D.J. Studholme et al., unpublished data). At the core of the system was a relational database (implemented in MySQL) that contained all previous sRNA sequencing data from our laboratory and all available data from other laboratories. As well as storing sRNA sequences, their abundances in each dataset and basic metadata about the biological origin of each sequencing library, the database also stored the results from all analyses performed on these sequences (matching against known miRNAs, matching against tRNA/rRNA, prediction of regulatory targets, match sites in plant and virus genomes). It also tabulated nucleotide frequencies at the 5’-end of the sRNAs [130]. The system included a set of Perl scripts for performing the analyses and populating the database. It also included a set of Perl CGI scripts that provided a web-based user interface that enabled the user to query the database and generate summary graphics, tables and statistics ‘on the fly’ and across all datasets. It was also closely integrated with a genome browser allowing sRNAs to be viewed in context with other genomic features.

Our sRNA database and analysis system was developed purely for use in house and has never been made public (though the source code can be provided on request to the author, and some of the code formed a starting point for the freely available UEA sRNA Toolkit). It is likely that other laboratories have developed similar in-house data management systems and do not have the resources to package these up into a form that can be easily deployed by other laboratories. However, a notable exception is the Arabidopsis Small RNA Project Database (ASRPDB) [180, 181]. This database began as an in-house system for the Carrington Laboratory at Oregon State University with aims similar to the system that we developed at The Sainsbury Laboratory, but it has been made publicly available. It has been given a professional quality user interface including several useful search tools and currently contains high-throughput sequencing results from the Carrington and Bartel laboratories. However, the author is aware of no publicly available comprehensive and up-to-date database offering search and analysis tools for all sRNA sequence data from all laboratories; the deepBase database [182] partially fulfils this role, but it has not been updated between November 2010 and November 2011. There is also at least one specialized databases dealing with plant microRNAs [183]. There is a pressing need for a comprehensive database of RNA sequences that is internationally supported for the long term [184], serving a similar function for RNA as which UniProt [185] fulfils for protein sequences.

**USEFUL TOOLS AND DATABASES FOR sRNA ANALYSIS ARE FREELY AVAILABLE**

All of the common steps in sRNA analysis workflows can be accomplished using freely available software tools (Table 2). One approach is to use a collection of tools from heterogeneous sources. The advantage of this is that it offers complete freedom to choose the most appropriate tool for each step. The steps may be linked together and automated using a scripting language. An alternative approach is to use a comprehensive one-stop-shop. The advantages include convenience of not having to source tools for each step and the reduced reliance on scripting and other computer-based skills that might be limiting for some laboratories. A good example of such an approach is the UEA sRNA Toolkit (Table 1). This offers most of the functionality that most biologists would need for analysing their sRNA sequencing data. It can be accessed via a webserver or the underlying software can be downloaded and installed and used locally. The standalone
tools are currently being substantially revised to improve performance and maintainability. The webserver is still available but will likely struggle to deliver as datasets increase in size and numbers of users increase. It is important to realize that many of the components of the UEA sRNA Toolkit are essentially high-level wrappers around existing tools for more low-level tasks. For example, it currently uses PatMan [186] to align sRNA sequences against a reference genome and uses the Vienna package [187] for predicting RNA secondary structures. However, it does also include some novel algorithms, such as FiRePat, for discovering patterns in expression profiles. Some alternatives that provide at least some of the same functionality as the UEA sRNA Toolkit include miRNAkey [188], miRanalyzer [189], SeqBuster [190], DSAP [191] and mirTools [192], though none of these are primarily intended for plant-derived data.

The miRNAkey software package [188] can perform most of the steps that would commonly be included in sequencing-based comparisons of miRNA profiles. It provides tools for removing adaptor sequences, identifying reads that match miRNAs documented in third-party databases (e.g. Rfam) and performing quantitative comparisons of each miRNA’s abundance between pairs of samples. The quantitative comparison includes normalization and testing for statistical significance. Furthermore, it provides a novel approach for dealing with cross-matching miRNA sequences. Although it does not offer the option to infer novel miRNAs from the data, miRNAkey does represent a comprehensive suite of utilities for complete comparative analysis of miRNA expression profiles. Complementary to miRNAkey is miRanalyzer [189], which identifies potential new miRNA sequences in large sequence data sets and compares

Table 1: The University of East Anglia sRNA Toolkit

<table>
<thead>
<tr>
<th>Tool</th>
<th>Function</th>
<th>Developed by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequence file pre-processing</td>
<td>Removes adapter sequences and converts from FastQ to FastA format.</td>
<td>S. Moxon</td>
</tr>
<tr>
<td>Filter</td>
<td>Removes sequences that match tRNA or rRNA turn-over products, fail to</td>
<td>F. Schwach</td>
</tr>
<tr>
<td>miRCat</td>
<td>Identifies candidate novel miRNA sequences.</td>
<td>S. Moxon</td>
</tr>
<tr>
<td>miRProf</td>
<td>Generates expression profile of known miRNAs.</td>
<td>F. Schwach</td>
</tr>
<tr>
<td>RNA hairpin folding and</td>
<td>Generates an annotated graphical representation of the secondary</td>
<td>F. Schwach</td>
</tr>
<tr>
<td>annotation</td>
<td>structure from a candidate miRNA precursor sequence.</td>
<td></td>
</tr>
<tr>
<td>FiRePat</td>
<td>Analyses time-course data to identify matches in expression profiles</td>
<td>I. Mohorianu</td>
</tr>
<tr>
<td>SiLoCo</td>
<td>Identifies sRNA-generative loci based on one or more datasets and</td>
<td>F. Schwach</td>
</tr>
<tr>
<td>ta-siRNA prediction</td>
<td>compares sets of loci between multiple datasets.</td>
<td></td>
</tr>
<tr>
<td>Plant target prediction</td>
<td>Identifies genomic loci that have significant phasing of aligned</td>
<td>S. Moxon</td>
</tr>
<tr>
<td></td>
<td>sRNAs and are therefore candidate ta-siRNA loci.</td>
<td></td>
</tr>
</tbody>
</table>

The UEA toolkit facilitates many of the tasks commonly encountered during analysis of data from large-scale sequencing of plant sRNA. The tools are available as standalone scripts that can be run locally on the user’s computer or can be run via the website at http://srna-tools.cmp.uea.ac.uk/plant/cgi-bin/srna-tools.cgi.

Table 2: Some useful web-based resources for analysing plant sRNA sequence data

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>UEA sRNA Toolkit</td>
<td>Comprehensive suite of tools for analysing large-scale sRNA sequencing data.</td>
<td><a href="http://srna-tools.cmp.uea.ac.uk/plant/cgi-bin/srna-tools.cgi">http://srna-tools.cmp.uea.ac.uk/plant/cgi-bin/srna-tools.cgi</a></td>
</tr>
<tr>
<td>ASRP DB</td>
<td>Database of Arabidopsis sRNA sequences.</td>
<td><a href="http://asrp.cgrb.oregonstate.edu/db/">http://asrp.cgrb.oregonstate.edu/db/</a></td>
</tr>
</tbody>
</table>

Studholme
their expression levels between samples. It can be configured for either plant- or animal-derived data.

THE PROBLEM OF DIFFERENTIAL EXPRESSION
Beyond discovering new species of sRNA, a common goal of high-throughput sequencing is to quantitatively compare expression profiles between different biological samples [172, 193]. In biomedical studies, microarrays are routinely used to simultaneously assay hundreds of known miRNAs quantitatively. Hybridization-based methods, including microarrays, have recently lost some ground to sequencing-based approaches, which have the advantage of a greater dynamic range and are not limited to assaying previously known sequences. The problems of quantitatively comparing expression levels of sRNAs are conceptually very similar to those of RNA-seq, that is, the sequencing of longer, protein-coding mRNA (or cDNA) [194]. For miRNAs, each genomic locus is represented by just one or two unique sequence tags. The number of sequence reads for each miRNA is simply taken as a proxy for its level expression. However, for more complex sRNA loci, the individual sRNA species have to be aggregated prior to counting; then the aggregated number of reads is taken as the proxy for expression level. The issues of quantification have recently been thoroughly discussed in the excellent review by [172]. However, it is worth pointing out here, that the much-touted advantages of the so-called next-generation sequencing technologies do not negate the need for adequate numbers of technical and biological replicates [195]. Furthermore, it should be noted that most methods for comparing expression profiles are only able to handle pair-wise comparisons between two treatments. Recently, a set-based method has been proposed that can handle comparisons of much more complex experimental designs [196].

REGULATORY TARGETS OF sRNAs CAN BE PREDICTED
Some sRNAs, notably miRNA and ta-siRNA, act as post-transcriptional regulators, binding to target transcripts in a sequence-specific manner and modifying RNA stability and/or rate of translation. For most known regulatory targets in plants, the sequence-specific binding depends on near perfect sequence complimentarity between the sRNA and its target. This led to the proposal of set of criteria for identifying likely target transcripts [197, 198] that has been used, with minor modifications, in numerous studies and tools [199–203]. There have also been attempts to integrate expression profiling with sequence-based cues to predict regulatory targets [204]. Predictions based on these rules have often been confirmed experimentally, e.g. using 5’-RACE with low levels of false positives. However, it is impossible to assess the rate of false negatives that arise due to the high stringency of these criteria.

NETWORKS OF sRNA INTERACTIONS MAY GENERATE EMERGENT PROPERTIES
The ability to predict the regulatory targets of regulatory sRNAs is useful for determining the functions of individual molecules, but it also has another important application—that is, the study of regulatory interactions at the systems level. For example, we recently modelled RNA-RNA interactions in Arabidopsis as a network of regulatory interactions based on empirical sRNA sequencing data and predictions of regulatory targets [205]. These network models suggested a system with emergent properties with high robustness to loss of elements. Further integrative systems level approaches [206] may hold the key to understanding the true significance of sRNA, the ‘dark-matter of genetics’ [207].

Key points
- Some miRNA and other sRNA can be predicted from genomic or EST sequence data.
- sRNAs can be directly detected by large-scale sequencing.
- Making sense of sRNA large-scale sequencing usually requires mapping to a reference genome.
- Different classes of sRNA give different patterns of alignment against the reference genome.
- Database systems are useful for managing large-scale sRNA sequencing data.

Acknowledgements
The author wishes to acknowledge the many useful discussions over several years with Frank Schwach and Simon Moxon.
References

44. Kurth Y, Takashi Y, Watanabe Y. The interaction between DCL1 and HY1L is important for efficient and


Deep sequencing of small RNAs in plants


