EMBnet.journal
Bioinformatics in Action

Olagunju et al. (2024) EMBnet.journal 29, e1037
http://dx.doi.org/10.14806/ej.29.0.1037

Research Papers

# Ds-Seq: An Integrated Pipeline for In Silico Small RNA Sequence Analysis for Host-pathogen Interaction Studies

**Temitayo Adebanji Olagunju[1,2], Angela Uche Makolo[2], Andreas Gisel[2,3]✉**

[1]University of Ibadan, Ibadan, Nigeria
[2]Bioscience Centre, International Institute of Tropical Agriculture (IITA), Ibadan, Nigeria
[3]Institute for Biomedical Technologies, National Research Council of Italy (CNR), Bari, Italy
Competing interests: TAO none; AUM none; AG none

**Abstract**

Plant-pathogen interactions activate molecular activities wherein the host defends the pathogen while the pathogen tries to suppress the plant response. Small RNAs (sRNAs) mediate major mechanisms, including post-transcriptional gene silencing, histone modification and DNA methylation by which plants respond to the presence of pathogens. Genome-wide profiling of host and pathogen sRNAs is therefore pivotal to uncovering the mechanisms underlying the host-pathogen interaction and mechanisms for host resistance. sRNA high throughput sequencing (HTS) data analysis often involves multiple stages/tools. Most necessary tools are accessible only through the command line, making it challenging for those without a high level of Unix/Linux skills. Furthermore, installation of some of these tools may become difficult due to dependencies and software version compatibility. We have developed an integrated open-source pipeline, Ds-Seq, for end-to-end in silico analysis of sRNA HTS data with improved reproducibility. The pipeline combines in-house scripts and public tools in a shell script, which can be invoked with a single command. The pipeline's usefulness has been demonstrated with testing on publicly available and published data from independent sRNA-seq datasets of host-pathogen interaction studies. Ds-Seq is available on GitHub, while a Docker image can be obtained from the Docker hub.

## Introduction

Small RNAs (sRNAs), generally between 20-30 nt long, have regulatory functions which are critical to many biological processes (Won *et al.*, 2014). The 21-24 nt subset, also referred to as micro RNAs (miRNAs), is evolutionarily conserved between species (Zhang *et al.*, 2006). In host-pathogen interaction, sRNAs play a significant role in the defense response of the host to pathogen invasion through the RNA interference (RNAi) machinery (Pumplin and Voinnet, 2013; Carbonell *et al.*, 2019). sRNAs can also regulate gene expression through DNA methylation and histone modification (Wang *et al.*, 2018; Tamiru *et al.*, 2018; Diezma-Navas *et al.*, 2019). These multiple pathways of sRNA-mediated gene expression regulation underscore the importance of genome-wide characterization of sRNAs as an essential first step to further investigation of molecular mechanisms orchestrated by sRNAs in response to different environmental cues.

High throughput sequencing (HTS) technologies such as small RNA sequencing (sRNA-seq) and computational analysis tools have been successfully applied to carry out genome-wide profiling of the sRNA landscape of many organisms to investigate roles played by sRNAs in different conditions (Li *et al.*, 2018; Hu *et al.*, 2020; Wenlei *et al.*, 2020).

Knowledge discovery using computational tools for in silico analysis of HTS data often involves multiple stages, determined mainly by the specific research questions to be answered. Each step usually entails selecting a computational tool from a buffet of available tools. Many of these tools are accessible only through a command-line interface (CLI) (Seemann, 2013), making it a challenge, especially for those not familiar with a Unix/Linux operating system environment (Xu *et al.*, 2014; Morais *et al.*, 2018; Joppich and Zimmer, 2019). The command-line design, however, favours a batch processing approach for computational pipelines where the output file of a stage, after manipulation, is parsed to the tool at the next step. Furthermore, installation of some of these required tools for analysis may become difficult due to dependencies and software version

**Article history**
Received: 23 September 2023
Accepted: 19 November 2023
Published: January 2024

compatibility with user operating systems (List *et al.*, 2017; Mangul *et al.*, 2018). More efficient and effective research can only be conducted when wet biologists are left to focus on knowledge discovery from computational analysis and not waste time setting up tools, troubleshooting or learning to use them (Smith, 2013; Smith, 2014). It is important to note that the problem of reproducibility of results (Stodden *et al.*, 2018; Beaulieu-Jones, 2017) also becomes pronounced against the backdrop of the challenges of bioinformatics tools installation and management.

Some pipelines for the analysis of sRNAs are already in existence but are not tailored for host-pathogen interaction experiments. The Java-based UEA sRNA Workbench (Beckers *et al.*, 2017) was developed to explore the sRNA landscape based mainly on the differential expression analysis of sRNA libraries. This tool is thus limited in its scope. sRNAPipe was developed by (Pogorelcnik, 2018) as a GUI-based pipeline on the Galaxy platform (Afgan *et al.*, 2018) for the exploratory analysis of sRNA libraries. The pipeline maps the reads to only the host genome assembly, after which they are classified into four main sub-groups for further exploration and does not include a module for differential expression analysis of the sRNAs in the libraries. sRNAbench and sRNAtoolbox 2019 (Aparico-puerta *et al.*, 2019), which is an update of the sRNAtoolbox (Rueda *et al.*, 2015), provides a more comprehensive suite of tools for the exploration of the sRNA landscape through a GUI, including differential expression analysis, visualisation of genome-mapped reads on a genome browser, and sRNA target prediction. Although the sRNAtoolbox provides a differential expression analysis module, unlike the UEA sRNA Workbench and the sRNAPipe, it does not offer segregation of the sRNAs by the library to aid quick identification of unique and common sRNAs in the libraries.

In this paper, we present Dockerized sRNA-Seq tool, Ds-Seq, an open-source in-silico analysis pipeline for investigating host-pathogen interaction in plants with available genome assemblies through profiling the sRNA landscape. With parameters specified by the user through a configuration file, it takes the fastq files of the sRNA-Seq libraries as input, maps the reads to the user-supplied host and pathogen genome assemblies, produces a differential expression profile of the sRNAs in the libraries, segregates the host-mapped sRNAs by the library, identifies known miRNAs, and predicts novel sRNAs. Only a single command is required to initiate the analysis with this pipeline, making it relatively easy to use in a CLI environment without requiring high technical skills in UNIX/Linux environment. We also introduced the use of a Docker container to provide a consistent environment for reproducibility of results (Baker and Penny, 2016) and eliminate the challenges occasioned by software versioning and compatibility issues. This pipeline has been tested with publicly available sRNA-Seq data from host-pathogen interaction studies with accompanying published results. The outcome of the testing generally showed agreement between results obtained by Ds-Seq and published data. Ds-Seq is freely available from the GitHub repository[1] and Docker hub with ID cephas/ds-seq[2].

## Materials, Methodologies and Techniques

Ds-Seq with (schematic shown in Figure 1) was developed with Perl, Python and R scripts, all wrapped in a single shell script that can be run with a single command. It is designed with a modular structure such that each module corresponds to a stage of the analysis that could be carried out separately with other bioinformatics tools. The user can select specific modules of interest or all the modules through a configuration file where all parameters for the analysis are defined. It is important to note that Ds-Seq does not contain a database, so all the files required for analysis, including reference genome sequences, have to be supplied by the user. Parameters for the independent tools are also to be defined by the user through an accompanying configuration file. Due to the wide variation of the sRNA family, based on the sequence length, the class of sRNA of interest to a user can be defined and controlled with the length parameters in the configuration file. Further exclusion of other types of non-coding RNAs such as tRNA, piwiRNAs, snoRNAs etc., can be made from the analysis through a multi-fasta file containing the sequences to be excluded. Sequences from Rfam (Kalvari *et al.*, 2020) are examples of sequences that could be excluded from any analysis carried out with Ds-Seq.

### Design

Ds-Seq is designed to analyze sRNA-seq data easily and with reproducibility, especially for wet biologists with minimal experience with Unix/Linux operating system command line interface. As such, the invocation of the pipeline is achieved with a single line of command after parameters have been defined in the accompanying configuration file and user files have been correctly placed. Each module of Ds-Seq has specific file requirements such that the modules selected by a user would determine the required input files and data. File requirements are shown in Table 1 for the modules (i) NGS sequences filtering, (ii) expression profiling, (iii) novel sRNA prediction, (iv) genome-wide host and pathogen sRNA mapping, and (v) conserved sRNA mapping. Reads mapping is carried out generally using Bowtie (Langmead, 2009), being a memory-efficient and ultra-fast short DNA sequence alignment program, using default parameters for mapping in addition to reporting the best alignment. At the same time, the user defines the number of accepted nucleotide mismatches between sequences through the configuration file.

[1]https://github.com/CEPHAS-01/small-RNASeq.ngs

[2]https://hub.docker.com/r/cephas/ds-seq

Olagunju et al. (2024) EMBnet.journal 29, e1037
http://dx.doi.org/10.14806/ej.29.0.1037



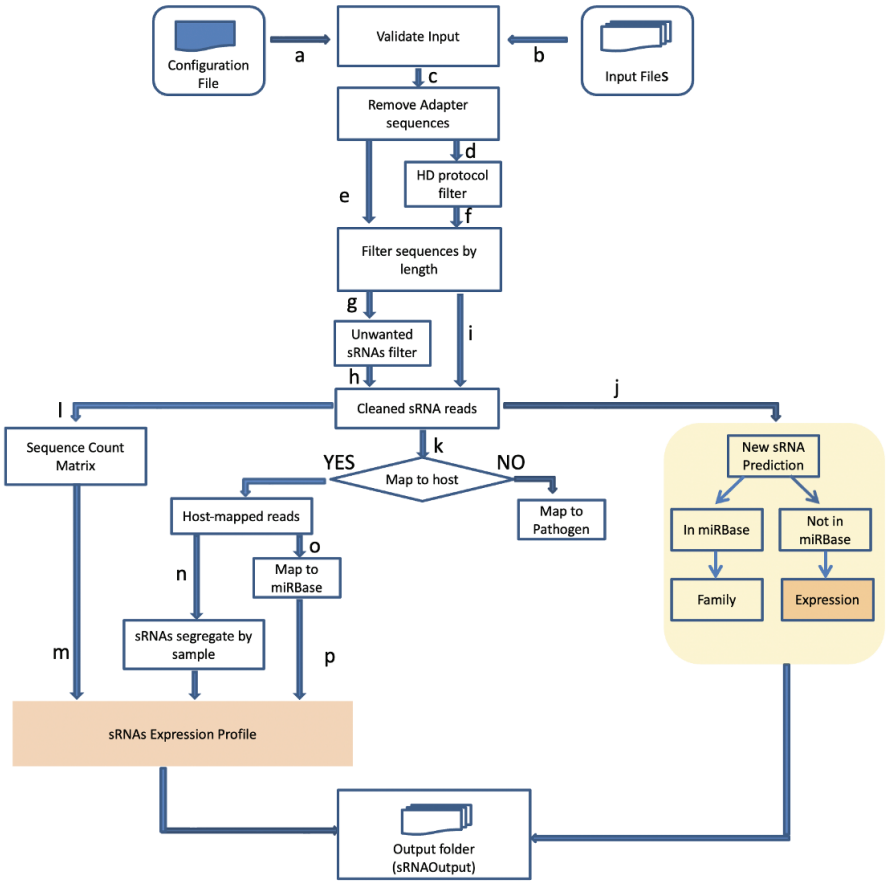


**Figure 1.** Schematic of sRNA sequencing analysis pipeline.
**(a)** Parameters and file paths are defined in a configuration file. **(b)** Raw sequence reads of the samples in .fastq or .gz file format, and other data are supplied to the validation module. **(c)** With successful validation stage, reads are passed on for removal of adapter sequences where raw reads without adapters are discarded while adapter sequences are removed from the other reads. **(d)** Adapter-cleaved reads are further processed if the HD protocol was used in the library preparation. **(e)** Adapter-cleaved reads are filtered for the length specified in the configuration file to produce clean reads. **(f)** Reads processed for HD protocol are filtered with length to produce clean reads. **(g)** Length-filtered reads are optionally processed to remove unwanted non-coding RNAs supplied by the user **(h)** Clean reads are produced for further analysis **(i)** cleaned reads are produced for further analysis **(j)** Cleaned reads are parsed for novel small RNA prediction. **(k)** Cleaned reads are mapped to the host genome and separated into host-mapped and unmapped sRNAs **(l)** Cleaned reads are parsed for expression matrix generation, and **(m)** expression matrices are parsed for differential expression analysis **(n)** Host-mapped sRNAs are segregated by sample **(o)** Host-mapped sRNAs are mapped to known miRNAs repository to identify known miRNAs **(p)** Known miRNAs expression profile in samples are produced.

Input raw sequence files can be left in compressed zip (.gz) format without the need to first decompress as the pipeline can discriminate the file type and handle it appropriately. The raw sequence files of the sRNA-Seq libraries must be saved in replicates in a folder named by the sample and located in the 'data' directory in a manner representative of the experimental design. A unique analysis ID is generated for each non-redundant sequence and is consistent across all the libraries. The checkpoint feature is implemented at strategic stages of the pipeline to avoid repeating steps previously executed, especially tasks requiring intensive compute time, such as reference sequence indexing. The configuration file can be used for processing other analyses while retaining the same parameters to enable the user to compare different libraries to identify similarities and differences.

The analysis modules in Ds-Seq are further explained in the following sections.

## NGS reads filtering

The NGS data filtering is carried out using in-house Perl scripts to remove adapter sequences and sequences without an adapter from the input raw reads library. An optional further clipping of 4 nt from the 5' and 3' ends of the cleaned reads is done if the High Definition protocol (Sorefan *et al.*, 2012) was used during the library preparation. For each input raw reads file, each sequence is aggregated into one, and its abundance in the input file is appended to the sequence header information to produce a multi-fasta file. Non-redundant reads with a length that meets the user-defined range and reads abundance not less than the defined threshold in the con-figuration file are retained for use downstream of the analysis pipeline as clean reads. An optional step

**Table 1.** Files requirement for each modular stage of the analysis pipeline.

| Analysis module | Other input parameters (through configuration file) | Reference genome assembly | Annotation file | Chromosome length file |
|---|---|---|---|---|
| **Filter NGS data** | Yes | No | No | No |
| **Host mapping** | Yes | Yes | No | No |
| **Conserved sRNA** | Yes | Yes | No | No |
| **Pathogen mapping** | Yes | Yes | No | No |
| **Novel sRNA prediction** | Yes | Yes | Yes | Yes |
| **Expression profile** | Yes | No | No | No |

also removes any other unwanted non-coding RNA sequences from the reads, such as sequences from Rfam (Kalvari *et al.*, 2020) or different user-defined sequences supplied through a multi-fasta file. A bar plot of the length distribution of the cleaned reads from all replicates of the samples is also produced to give an insight into the profile of the sRNAs within the libraries and as a further check of the quality of the sRNA-Seq libraries. This information helps determine the quality of the library. It could also provide information on the specific molecular pathway origin of the sRNAs based on the activities of Dicer (Mueth *et al.*, 2015).

### Differential Expression Profiling

This module in the pipeline is used to reveal the differential expression profile of all sRNAs across the libraries and experimental conditions of interest. Count matrices of the cleaned reads across pairwise combinations of all libraries are produced, using the sequence as a unique identifier to achieve sequence identity uniformity across all the libraries. These matrices are parsed as input into the edgeR (Robinson *et al.*, 2009; McCarthy *et al.*, 2012) package for the differential expression analysis. Reads that meet the fold change and p-value criteria defined in the configuration file are returned as differentially expressed.

### Conserved/known miRNAs identification

Micro RNAs (miRNAs) belong to a sub-class of sRNAs with sequence lengths between 21-24 nt. Some miRNAs are evolutionarily conserved across species (Zhang *et al.*, 2006) and are identified through a homology search of the known mature sequences in the miRNA repository miRBase (Kozomara *et al.*, 2019) (www.mirbase.org). In this pipeline module, all known mature sequences hosted in miRBase are supplied as a multi-fasta sequence file indexed with Bowtie (Langmead *et al.*, 2009) before mapping. The user can alternatively provide a list of conserved sequences of interest in a multi-fasta format as a reference in place of the mature sequences from miRBase. All sequences with the best hit within the defined acceptable nucleotide mismatch parameter defined in the configuration file mapping to this repository are reported as known sRNAs. The sequences specific to the host are reported as known host sRNAs.

### Prediction of novel sRNAs

Plants produce stress-responsive sRNAs in response to biotic and abiotic stressors. Due to the specificity of production of these sRNAs (Sunkar *et al.*, 2012), novel prediction is required to uncover those present in the sRNA-Seq libraries. To this aim we use miRDeepP (Yang and Li, 2011), an open-source software designed specifically for predicting sRNAs in plants. mirDeepP comprises nine different Perl scripts. The critical parameters required at some of the stages of this analysis have been included in the parameters supplied by the user through the configuration file. The reader is referred to (Yang and Li, 2011) for further information on the distinct stages, and the parameters required. The predicted sRNA sequences are further filtered by mapping to miRBase to discard known miRNAs, while unmapped sequences are retained as potential novel sRNAs.

### Genome-wide Host and Pathogen sRNA Profiling

To differentiate between host and pathogen(s) sRNAs, the raw reads are mapped against the corresponding reference genomes supplied by the user. The indexing of the genomes and the subsequent mapping is done by the mapping software and does not require any further action by the user. The sequences are mapped first to the host reference genome, and only reads that fail to map to the reference genome are mapped to the pathogen reference genome. This step provides the mapping profiles of each sRNA-Seq library, reporting the number of reads mapped to the reference genomes (host and pathogen), the distribution of the sequence lengths of interest defined by the user in the configuration file across all the libraries, and the loci of each sRNA on the reference genome. A repository of pathogens genome assemblies can be supplied as a reference in multi-fasta file format for a comprehensive pathogen sRNA profiling.

## Input

To use Ds-Seq, it is expected that the NGS sequences have passed quality control and are deemed suitable for further downstream analysis. The input files to the pipeline consist mainly of (i) a configuration text file and (ii) data, including the raw single-ended sRNA sequence reads in fastq or zipped fastq formats (.gz),

**Table 2.** List of third-party tools used in the pipeline and their sources.

| Usage | Tool | Source |
|---|---|---|
| Differential Expression | EdgeR | https://bioconductor.org/packages/release/bioc/html/edgeR.html |
| Sequence Alignment | Bowtie | http://bowtie-bio.sourceforge.net/index.shtml |
| Novel miRNA Prediction | miRDeep-P | http://sourceforge.net/projects/mirdp/ |
| MFE: Vienna RNA Package (RNA fold) | RNAfold | https://www.tbi.univie.ac.at/RNA/ |
| Plots | GGPlot2 | https://cran.r-project.org/web/packages/ggplot2/index.html |

annotation file corresponding to the genome assembly build, chromosome length file and reference sequences (Table 2). The pipeline can parse unzipped input fastq.gz files without requiring any pre-processing. A validation module checks and validates all the input parameters from the configuration and input files. If an error is flagged, the analysis does not commence, and a report is written to a log file.

## Output

The results of the different stages of the analysis are organised into subfolders based on the modules of the pipeline, and all the sub-folders are presented in a single output folder. A schematic showing the organisation of the pipeline output is shown in Supplementary Figure S1[3], while a more detailed description is presented in Supplementary Table 13[4]. The output files are formatted as tab-delimited flat files to make it easy to port the results to other tools for further downstream analysis. Results produced include the profile of sRNA in the samples high-lighting the host and pathogen sequences, the differential expression profile between libraries, conserved sRNAs, and novel predicted sRNAs. Publication-ready plots of the sRNA length distribution across libraries and frequency of the nucleotides at the 5' positions of the reads are produced. For differential expression analysis, volcano plots, a multidimensional scaling plot (MDS) and a biological coefficient of variation (BCV) (McCarthy *et al.*, 2012) plot are also produced.

## Validation

Ds-Seq was validated using two publicly available datasets obtained from host-pathogen interaction studies with published results.

(i) Study A: sRNA-seq data obtained from the work of (Yang *et al.*, 2018) where the regulatory roles of Rice Stripe Virus (RSV)-derived small interfering RNAs were investigated in rice. Although this work entailed comparative sRNA data analysis from rice and the insect vector Laodelphax striatellus, we focused on only the RSV interaction with rice. Two samples of mock and RSV-infected rice in two replicates each were used. The data was obtained from the National Center for Biotechnology Information (NCBI), accession number GSE113555.

(ii) Study B: NGS data obtained in a study of sRNAs from the interaction between Turnip Mosaic Virus (TuMV) and the Tanto and Drakkar cultivars of oilseed rape Brassica napus (Pitzalis *et al.*, 2020). Only the data from the Drakkar cultivar was used for testing Ds-Seq. The sRNA-seq data in this study were retrieved from the NCBI Sequence Read Archive (SRA) using accession code PRJNA508739.

## Pipeline Availability

Ds-Seq can be downloaded as scripts from the GitHub repository[1] and run on a UNIX/Linux operating system or obtained as a docker image from Docker hub with the ID cephas/ds-seq[2]. Further details of how to download and use the pipeline are contained in the README.md file on the GitHub repo. A user manual describing how to use the pipeline is available for download at the Github repository.

## Technical Information

The versions of the software used in the Docker container for this pipeline and under which it has been tested are listed as follows:

1. Ubuntu - 18.04
2. R-base - 3.6.1
3. Perl - 5.26.1
4. Python - 2.7.17
5. RNAfold 2.4.11

Information on third-party tools used in the pipeline is presented in Table 2. We recommend running the pipeline using the Docker container since all the software are already bundled with the image.

However, the pipeline can still be run as scripts on a UNIX/Linux machine if an environment with these versions of tools is provided. This is not to say that other versions of these software will not work, but that the pipeline has performed correctly in an environment with these specific software versions.

# Supporting data

The two published datasets used for testing the pipeline were obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database for Study A using accession GSE113555, and Sequence Read Archive (SRA) for Study B using accession PRJNA508739. All the data produced

[3]http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/1037/1037_supp_fig

[4]http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/1037/1037_supp_tab

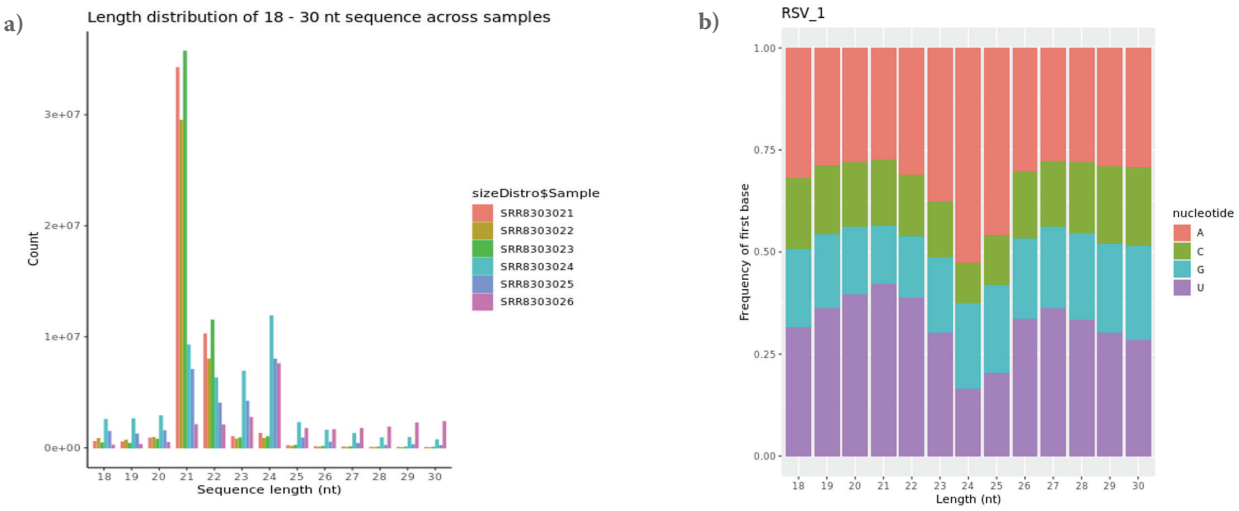


**Figure 2.** Examples of plots generated by Ds-Seq showing **(a)** the length distribution of the Mock and TuMV-infected Brassica napus sRNA reads across all libraries showing higher accumulation of 21-24 nt se-quences over other lengths in Study A. **(b)** Nucleotide frequency distribution for the sRNA sequence lengths in the two Rice RSV-infected samples showing dominance of nucleotide U at 21 nt and A at 24 nt positions for Study B.

by the pipeline have been reported in this paper and the supplementary files. A test data for the pipeline will be made available for review on request pending the time that it would be hosted permanently in a public repository.

## Results

Ds-Seq was tested with two different publicly available datasets , as highlighted in the validation section. For each test, the versions of the public dataset and the parameters used for the analysis by the authors in their study were replicated as much as possible in the pipeline.

For Study A (Yang *et al.*, 2018), the average cleaned reads produced with Ds-Seq for the Mock and RSV-treated samples were 13,664,509 and 15,206,600, respectively, compared to 13,876,277 and 15,400,564 reads reported in the study for the same data (Supplementary Table 1[4]). Furthermore, the average percentage of host-mapped reads obtained with Ds-Seq was 69.92% for Mock samples and 66.73% for RSV-treated samples while in contrast, 64.90% and 59.92% were reported respectively for Mock and RSV samples in Study A (Supplementary Table 2[4]). Negligible reads were reportedly mapped to the RSV genome for Mock samples (0.00%) by Ds-Seq and in Study A, while for RSV-treated samples 0.54% and 0.29% reads were reported mapped to RSV genome by Ds-Seq and in Study A respectively (Supplementary Table 2[4]).

The sRNA reads length distribution produced by Ds-Seq showed an accumulation of 21-24 nt sRNAs over other reads length (Supplementary Figure S3[3]), in agreement with the report in Study A. The distribution of the nucleotides at each position of the reads in the libraries produced by the pipeline revealed that at the 5' end, most of the 21nt sRNAs had Uracil (U) while most of the 24 nt sRNAs had Adenosine (A) (Figure 1B and Supplementary Figure S2[3]), consistent with the report in Study A. Differential expression profiles upon virus infection as reported in the study showed that 450 and 1,558 sRNAs were upregulated and downregulated respectively (Yang *et al.*, 2018), compared with 3789 and 1924 sRNAs identified by Ds-Seq at $|\log FC| \geq 2.0$ and adjusted p-value < 0.05 (Supplementary Tables 7[4] and 8[4]). Although not reported in Study A, Ds-Seq identified 628 (167), 687 (172), 925 (178) and 749 (176) conserved miRNAs (Rice-specific conserved miRNAs) in the Mock1, Mock2, RSV1 and RSV2 samples respectively (Supplementary Table 1[4]). Ds-Seq also reported 81 and 115 sRNAs from the Mock and RSV samples, respectively, as likely novel sRNAs predicted by miRDeep-P (Supplementary Tables 9[4] and 10[4]), which were also not mentioned in Study A.

In Study B (Pitzalis *et al.*, 2020), the percentage of cleaned reads produced by Ds-Seq was between 48.2% to 74.8% for the Mock samples and 83.5% and 95.4% for the infected samples (Supplementary Table 5[4]). In comparison, 48.3% to 74.5% (Supplementary Table 5[4]) and 83.8% to 96.0% were reported in Study B for the Mock and infected samples respectively. For clean reads mapped to the Drakkar transcriptome, Ds-Seq re-ported 48.4% to 51.9% in the Mock samples, and 19.7% to 19.9% in the infected samples, as well as 0.01% to 0.03% and 30.7% to 33.6% of the reads, mapped to the virus genome in the Mock and virus-infected samples respectively (Supplementary Table 6[4]). This is compared with Study B, where 41.9% to 48.6% of the Mock samples mapped to the Drakkar (host) transcriptome and 13.4% to 15.3% mapped to the virus genome (Sup-plementary Table 6). Similarly, 0.01% to 0.06% of the cleaned reads mapped to the virus genome in the Mock sample, while in the infected samples, 64.56% to 68.23% of the reads mapped to the virus (Supplementary Table 6[4]). The sRNA profile

Olagunju et al. (2024) EMBnet.journal 29, e1037
http://dx.doi.org/10.14806/ej.29.0.1037



**Table 3.** Comparison of bioinformatics tools and parameters used at specific stages of the analysis by Ds-Seq and the independent investigation of the published datasets used to test Ds-Seq.

| Analysis | Adapter removal | Reads mapping (parameters) | Differential expression (parameters) | Conserved miRNAs reference |
|---|---|---|---|---|
| Ds-Seq | In-house Perl script | Bowtie (Langmead *et al.*, 2009) (-a --best)* | edgeR ( Robinson *et al.*, 2009; McCarthy *et al.*, 2012) (p-value < 0.05) | mirBase (Kozomara *et al.*, 2019) |
| Study A ( Yang *et al.*, 2018) | Cutadapt ( Martin, 2011) | Bowtie ( Langmead *et al.*, 2009)) | edgeR ( Robinson *et al.*, 2009; McCarthy *et al.*, 2012) (adjusted p-value < 0.05) | |
| Study B ( Pitzalis *et al.*, 2020) | Cutadapt ( Martin, 2011) | Bowtie2 ( Langmead and Salzberg, 2012)) (-t -N 0–end- to-end–very-sensitive–score-min C,0,0) | DESeq2 (Love *et al.*, 2014) ( >= 150 mean reads and p-value < 0.05) | mirBase ( Kozomara *et al.*, 2019) and other sources described in the publication. |

generated by Ds-Seq showed a general accumulation of the 21-24nt reads more than other lengths. Still, more 24 nt reads were recorded in the mock-treated samples while 21-22 nt reads were the primary sizes in the infected plants (Figure 2A), as was reported in Study B.

The number of differentially expressed sRNA reads was reported from Ds-Seq as 49,426 and 18,255 upregulated and downregulated, respectively, at |logFC| ≥ 1.5 and pValue < 0.05 (Supplementary Tables 11[4] and 12[4]). In Study B, differential expression was reported for only conserved miRNAs present in the samples with at least ten mean normalized reads showing that about 150 miRNAs were upregulated and 90 were downregulated (Pitzalis *et al.*, 2020). For the conserved miRNAs from miRBase Ds-Seq reported between 575 and 1302 conserved miRNAs compared to 1047 identified in Study B, while those specific to B. napus ranged from 64 to 73 miRNAs across both mock and infected samples (Supplementary Table 5[4]). Some plots from Ds-Seq from the analysis of data from Study A and Study B are shown in Figure 2.

## Discussion

sRNA profiling has emerged as an essential step in studies regarding the control of gene expression in plants due to the implication of sRNAs in various molecular mechanisms by which the growth and develop-ment of plant species are regulated. These multiple mechanisms underscore the importance of a pipeline for sRNA profiling as a prime step toward discovering the molecular underpinnings of gene regulation involving sRNAs. In host-pathogen interaction, sRNAs play a significant role in the defense response of the host to pathogen invasion. The profiling of the sRNA landscape in host-pathogen interaction using NGS data opens a path toward the identification of sRNAs involved in the host defense and to further elucidates the molecular mechanism of the host defense response or pathogen action (Yang *et al.*, 2018; Pitzalis *et al.*, 2020). sRNAs of interest can then be extracted for further downstream analysis, such as identification of the target genes, co-regulation or DNA methylation.

In this paper, we introduce Ds-Seq, an integrated in silico pipeline for sRNA studies in plant host-pathogen interaction. The pipeline performance and utility were tested on two publicly available NGS data sets from which independent genome-wide sRNA analysis results have been published. The results of testing the automated pipeline generally showed agreement with the reported results of the independent analysis of the two published datasets (Yang *et al.*, 2018; Pitzalis *et al.*, 2020).

Ds-Seq identified a lower number of host-specific conserved miRNAs than the reported figures in the independent analysis, which could be attributed to the slight differences in tools with the runtime options applied at specific stages of the analyses.

In fact, two factors were noted regarding the difference with the results of the analysis of the conserved miRNA in Study B. Firstly, reads mapping was done using Bowtie2 (Langmead and Salzberg, 2012) with further restrictive parameters to filter the output, compared to the use of Bowtie (Langmead *et al.*, 2009) with all alignments reported in the pipeline (Table 3). Secondly, in Ds-Seq, only B. napus was used as the conserved miRNAs of interest from miRbase. In contrast, in the analysis of (Pitzalis *et al.*, 2020), the pool of conserved miRNAs was drawn from B. napus, B. rapa, B. oleracea, A. lyrate and A. thaliana (Table 3). The configuration file used for Ds-Seq can take only one plant name as an argument for the conserved miRNA of interest and could not capture the multiple plant names specified in Study B . This would be addressed in future updates to the pipeline to permit specifying multiple names of plants of interest.

A difference was also recorded between the number of differentially expressed sRNAs in study A and the results from the pipeline. A total of 384 and 181 sRNAs were reported to be upregulated and downreg-ulated respectively by the pipeline using the parameter p-value < 0.05. This was different from the reported values in the independent study, where differentially expressed sRNAs

were defined using an adjusted p-value < 0.05. A further step was taken on the results from the pipeline to extract differentially expressed sRNAs using adjusted p-value as was used in the independent study of Yang *et al.* (2018) to have a common basis for comparison. The result still showed a disparity in the pipeline-reported values and the independent analysis results, which could be due to omission in reporting some steps in the analysis or some parameters used in the separate study published by the authors.

Furthermore, a high number of differentially expressed sRNAs were reported by Ds-Seq with the data from study B. EdgeR (Robinson, 2009; McCarthy *et al.*, 2012) was used in the pipeline and differentially expressed reads were defined by the user through the configuration file using a fold change cut-off and p-value. In Pitzalis *et al.* (2020), however, DESeq2 (Love *et al.*, 2014) was used with extra parameters based on the number of reads used to filter the results. Therefore, using different tools (Maza, 2016; Costa-Silva *et al.*, 2017) could be responsible for the observed disparity in reported differentially expressed sRNAs.

The slight differences in the observed results will be addressed in the future release of Ds-Seq, especially those due to the use of different tools. We have plans to include more tools at various stages of the analysis to present users with options to choose from.

For differential expression analysis, the pipeline has been tested successfully on samples with two and three biological replicates each but can support samples with more than three biological replicates. The next iteration of Ds-Seq will accommodate other alignment software for the reads mapping module and integration of a module to predict the gene targets of the sRNAs.

Importantly, to ensure the reproducibility of the results (Baker and Penny, 2016), the pipeline can be containerized from its Docker image obtainable from the Docker hub repository. Application containerization with Docker[5] is employed to provide a consistent and self-contained software environment for running applications. Only the required dependencies of an application are installed before deployment to a host operating system. Docker containers have been demonstrated to promote reproducibility of genomic data analysis with minimal adverse effects on the outcome (Di Tommaso *et al.*, 2015). Aside from ensuring a consistent environment for applications, this approach also eliminates software version compatibility issues in application deployment.

## Conclusions

In this paper, we have presented Ds-Seq, an automated pipeline for in silico analysis of sRNAs which can be run from a UNIX/Linux machine command line interface or as a Docker container. The pipeline combines several open-source sRNA analysis tools, in-house Perl, Python and R scripts, into a single shell script that runs end-to-end sRNA analysis from sRNA-Seq libraries with a single command. Although Ds-Seq was designed and applied to the study of plant host-pathogen interaction, it can also be adapted to other studies that seek to investigate the influence of abiotic factors on any plant with an available genome assembly. We have demonstrated the pipeline's capabilities using two publicly available datasets, and the results obtained generally indicated agreement with the published results from the same datasets, although with some differences attributable to the use of different tools and analysis parameters. The observed differences between the published dataset and the results of this pipeline underscore the challenge of reproducibility of analysis results which can be eliminated with the use of a predefined analysis pipeline such as Ds-Seq in a Docker environment.

### Key Points

- An in-silico automated end-to-end analysis pipeline, Ds-Seq, is presented for small RNA profiling from NGS data in a host-pathogen interaction scenario.
- The pipeline has been tested with publicly available published datasets from host-pathogen interaction studies and the results showed general agreement with those obtained from the independent analysis of the datasets used.
- Ds-Seq containerization with Docker image promotes reproducibility of results through a consistent software environment.
- The pipeline features a modular structure that enables a user to choose modules of interest through a configuration file.
- Legible reports and results are presented as tab-delimited flat files for portability to other tools for further downstream analysis and publication-ready plots.

## Funding

This work was supported by the Bill and Melinda Gates Foundation through the grant INV-008053 "Metabolic Engineering of Carbon Pathways to Enhance Yield of Root and Tuber Crops" and Roots and Tubers Pro-gramme (RTB) of the Consultative Group on International Agricultural Research (CGIAR).

Data Availability Statement: The two published datasets used for testing the pipeline were obtained from the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) database for Study A using accession GSE113555, and Sequence Read Archive (SRA) for Study B using accession PRJNA508739. All the data produced by the pipeline have been reported in this paper and the supplementary files.

A test data for the pipeline will be made available for review on request pending the time that it would be hosted permanently in a public repository.

Acknowledgments: The authors are grateful to Yinka Fakunle and his ICT team for their help in maintaining the bioinformatics infrastructure.

## References

Baker,M. and Penny,D. (2016) Is there a reproducibility crisis? Nature, **533**, 452–454. http://dx.doi.org/10.1038/d41586-019-00067-3

[5]https://www.docker.com

Beaulieu-Jones,B.K. and Greene,C.S. (2017) Reproducibility of computational workflows is automated using continuous analysis. Nat. Biotechnol., **35**, 342–346. http://dx.doi/10.1038/nbt.3780

Carbonell,A., Lisón,P. and Daròs,J.A. (2019) Multi-targeting of viral RNAs with synthetic trans-acting small interfering RNAs enhances plant antiviral resistance. Plant J., **100**, 720–737 http://dx.doi.org/10.1111/tpj.14466

Costa-Silva,J., Domingues,D. and Lopes,F.M. (2017) RNA-Seq differential expression analysis: An ex-tended review and a software tool. PLoS One, **12**, 1–18. http://dx.doi.org/10.1371/journal.pone.0190152

Diezma-Navas,L., Pérez-González,A., Artaza,H., Alonso,L., Caro,E., Llave,C. and Ruiz-Ferrer,V. (2019) Crosstalk between epigenetic silencing and infection by tobacco rattle virus in Arabidopsis. Mol. Plant Pathol., **20**, 1439–1452. http://dx.doi.org/10.1111/mpp.12850

Di Tommaso,P., Palumbo,E., Chatzou,M., Prieto,P., Heuer,M.L. and Notredame,C. (2015) The impact of Docker containers on the performance of genomic pipelines. PeerJ, 2015, 1–10. http://dx.doi.org/ 10.7717/peerj.1273

Hu,T., Wang,T., Li,H., Wassie,M., Xu,H. and Chen,L. (2020) Genome-wide small RNA profiling reveals tiller development in tall fescue (Festuca arundinacea Schreb). BMC Genomics, **21**, 1–13. http://dx.doi.org/10.1186/s12864-020-07103-x

Joppich,M. and Zimmer,R. (2019) From command-line bioinformatics to bioGUI. PeerJ, 2019. http://dx.doi.org/10.7717/peerj.8111

Kozomara,A., Birgaoanu,M. and Griffiths-Jones,S. (2019) MiRBase: From microRNA sequences to function. Nucleic Acids Res., **47**, D155–D162 http://dx.doi.org/10.1093/nar/gky1141

Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol., **10**. http://dx.doi.org/10.1186/gb-2009-10-3-r25

Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. Nat. Methods, **9**, 357–359. http://dx.doi.org/10.1038/nmeth.1923

Li,A., Li,G., Zhao,Y., Meng,Z., Zhao,M., Li,C., Zhang,Y., Li,P., Ma,C. Le, Xia,H., *et al.* (2018) Combined small RNA and gene expression analysis revealed roles of miRNAs in maize response to rice black-streaked dwarf virus infection. Sci. Rep., **8**, 1–14. http://dx.doi.org/10.1038/s41598-018-31919-z

List,M., Ebert,P. and Albrecht,F. (2017) Ten Simple Rules for Developing Usable Software in Computa-tional Biology. PLoS Comput. Biol., **13**, 1–5. http://dx.doi.org/10.1371/journal.pcbi.1005265

Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol., **15**, 550. http://dx.doi.org/10.1186/s13059-014-0550-8

Mangul,S., Mosqueiro,T., Abdill,R.J., Duong,D., Mitchell,K., Sarwal,V., Hill,B., Brito,J., Littman,R.J., Statz,B., *et al.* (2018) Challenges and recommendations to improve installability and archival stability of omics computational tools. bioRxiv. http://dx.doi.org/10.1101/452532.

Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMB-net.journal, **17**, 10–12. http://dx.doi.org/10.14806/ej.17.1.200

Maza,E. (2016) In papyro comparison of TMM (edgeR), RLE (DESeq2), and MRN normalization methods for a simple two-conditions-without-replicates RNA-seq experimental design. Front. Genet., **7**, 1–8. http://dx.doi.org/10.3389/fgene.2016.00164

Morais,D., Roesch,L.F.W., Redmile-Gordon,M., Santos,F.G., Baldrian,P., Andreote,F.D. and Pylro,V.S. (2018) BTW-Bioinformatics Through Windows: An easy-to-install package to analyze marker gene data. PeerJ, 2018, 10–16. http://dx.doi.org/10.7717/peerj.5299

McCarthy,D.J., Chen,Y. and Smyth,G.K. (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res., **40**, 4288–4297. http://dx.doi.org/10.1093/nar/gks042

Mueth,N.A., Ramachandran,S.R. and Hulbert,S.H. (2015) Small RNAs from the wheat stripe rust fungus (Puccinia striiformis f.sp. tritici). BMC Genomics, **16**, 1–16. http://dx.doi.org/10.1186/s12864-015-1895-4

Pitzalis,N., Amari,K., Graindorge,S., Pflieger,D., Donaire,L., Wassenegger,M., Llave,C. and Heinlein,M. (2020) Turnip mosaic virus in oilseed rape activates networks of sRNA-mediated interactions between viral and host genomes. Commun. Biol., **3**, 1–16. http://dx.doi.org/10.1038/s42003-020-01425-y

Pumplin,N. and Voinnet,O. (2013) RNA silencing suppression by plant pathogens: Defence, coun-ter-defence and counter-counter-defence. Nat. Rev. Microbiol., **11**, 745–760. http://dx.doi.org/10.1038/nrmicro3120

Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2009) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, **26**, 139–140. http://dx.doi.org/ 10.1093/bioinformatics/btp616

Seemann,T. (2013) Ten recommendations for creating usable bioinformatics command line software. Gi-gascience, **2**, 2–4. http://dx.doi.org/0.1186/2047-217X-2-15

Smith,D.R. (2013) The battle for user-friendly bioinformatics. Front. Genet., **4**, 1–2. http://dx.doi.org/10.3389/fgene.2013.00187

Smith,D.R. (2014) Buying in to bioinformatics: An introduction to commercial sequence analysis software. Brief. Bioinform., **16**, 700–709. http://dx.doi.org/10.1093/bib/bbu030

Sorefan,K., Pais,H., Hall,A.E., Kozomara,A., Griffiths-Jones,S., Moulton,V. and Dalmay,T. (2012) Re-ducing ligation bias of small RNAs in libraries for next generation sequencing. Silence, **3**, 1–11. http://dx.doi.org/10.1186/1758-907X-3-4

Stodden,V., Seiler,J. and Ma,Z. (2018) An empirical analysis of journal policy effectiveness for compu-tational reproducibility. Proc. Natl. Acad. Sci. U. S. A., **115**, 2584–2589. http://dx.doi.org/ 10.1073/pnas.1708290115.

Sunkar,R., Li,Y.F. and Jagadeeswaran,G. (2012) Functions of microRNAs in plant stress responses. Trends Plant Sci., **17**, 196–203. http://dx.doi.org/10.1016/j.tplants.2012.01.010

Yang,X. and Li,L. (2011) miRDeep-P: A computational tool for analyzing the microRNA transcriptome in plants. Bioinformatics, **27**, 2614–2615. http://dx.doi.org/10.1093/bioinformatics/btr430

Tamiru,M., Hardcastle,T.J. and Lewsey,M.G. (2018) Regulation of genome-wide DNA methylation by mobile small RNAs. New Phytol., **217**, 540–546. http://dx.doi.org/10.1111/nph.14874

Wang,C., Wang,C., Xu,W., Zou,J., Qiu,Y., Kong,J., Yang,Y., Zhang,B. and Zhu,S. (2018) Epigenetic changes in the regulation of Nicotiana tabacum response to cucumber mosaic virus infection and symptom recovery through single-base resolution methylomes. Viruses, **29**;10(8):402. http://dx.doi.org/ 10.3390/v10080402

Wenlei,C., Xinxin,C., Jianhua,Z., Zhaoyang,Z., Zhiming,F., Shouqiang,O. and Shimin,Z. (2020) Com-prehensive Characteristics of MicroRNA Expression Profile Conferring to Rhizoctonia solani in Rice. Rice Sci., **27**, 101–112. http://dx.doi.org/10.1016/j.rsci.2019.04.007

Xu,G., Strong,M.J., Lacey,M.R., Baribault,C., Flemington,E.K. and Taylor,C.M. (2014) RNA CoMPASS: A dual approach for pathogen and host. PLoS One, **9**. http://dx.doi.org/10.1371/journal.pone.0089445

Yang,M., Xu,Z., Zhao,W., Liu,Q., Li,Q., Lu,L., Liu,R., Zhang,X. and Cui,F. (2018) Rice stripe vi-rus-derived siRNAs play different regulatory roles in rice and in the insect vector Laodelphax striatellus. BMC Plant Biol., **18**(1):219. http://dx.doi.org/10.1186/s12870-018-1438-7

Zhang,B., Pan,X., Cannon,C.H., Cobb,G.P. and Anderson,T.A. (2006) Conservation and divergence of plant microRNA genes. Plant J., **46**(2):243-59. http://dx.doi.org/10.1111/j.1365-313X.2006.02697.x