# Automated analysis of small RNA datasets with RAPID

Sivarajan Karunanithi[1,2,4], Martin Simon[3,6], Marcel H. Schulz[1,5]

February 20, 2019

1 Cluster of Excellence for Multimodal Computing and Interaction, Saarland University and Department for Computational Biology & Applied Algorithmics, Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany.
2 Graduate School of Computer Science, Saarland University, Saarland Informatics Campus, Saarbrücken, Germany.
3 Molecular Cell Dynamics Saarland University, Centre for Human and Molecular Biology, Saarland University, Saarbrücken, Germany.
4 Institute for Cardiovascular Regeneration, Goethe-University Hospital, Theodor Stern Kai 7, Frankfurt, Germany.
5 German Centre for Cardiovascular Research (DZHK), Partner site RheinMain, Frankfurt, Germany.
6 Molecular Cell Biology and Microbiology, Wuppertal University, Gausstraße 20, Wuppertal, Germany.

### Abstract

**Summary:** Understanding the role of short-interfering RNA (siRNA) in diverse biological processes is of current interest and often approached through small RNA sequencing. However, analysis of these datasets is difficult due to the complexity of biological RNA processing pathways, which differ between species. Several properties like strand specificity, length distribution, and distribution of soft-clipped bases are few parameters known to guide researchers in understanding the role of siRNAs. We present RAPID, a generic eukaryotic siRNA analysis pipeline, which captures information inherent in the datasets and automatically produces numerous visualizations as user-friendly HTML reports, covering multiple categories required for siRNA analysis. RAPID also facilitates an automated comparison of multiple datasets, with one of the normalization techniques dedicated for siRNA knockdown analysis, and integrates differential expression analysis using DESeq2. RAPID is available under MIT license at https://github.com/SchulzLab/RAPID.We recommend using it as a conda environment available from https://anaconda.org/bioconda/rapid.

## Introduction

Widespread availability of small RNA (sRNA) sequencing technologies drives the biological community in unraveling the pivotal role of sRNA molecules. Micro RNA (miRNA), short interfering RNA (siRNA), piwi-interacting RNA (piRNA), small nucleolar RNA (snoRNA), and trans-acting RNA (taRNA) are some members of the sRNA family. In a wide range of organisms, these sRNA molecules play crucial roles in gene regulation [3]. Although, miRNAs are the most widely studied sRNA molecules, a growing interest can be seen in other sRNA classes, like siRNAs. With improved mechanistic understanding of siRNA function, siRNAs are increasingly used as therapeutic agents in drug discovery [23]. Using siRNAs in therapy requires a solid understanding of siRNA biogenesis, and behavior.

Understanding siRNA biogenesis and function often involves the computation of various sequence properties like length, strand of origin, and soft-clipped nucleotides of sRNA molecules. A myriad of available sRNA analysis tools substantiate the complexity in analyzing sRNA data sets. Existing sRNA analysis tools can be broadly categorized into two categories based on their function. (i) Tools which are dedicated to predict novel miRNAs, piRNAs, etc. using diverse computational strategies. This list includes methods such as Shortstack [1], miRDeep2 [9], iMir [10], Piano [32], etc. (ii) The secondary focus of many sRNA analysis tools is to annotate, and perform Gene Ontology (GO) enrichment analysis of known, or predicted sRNAs. Examples include miRTools2 [33], iSmart [22], and CPSS [31]. However, such annotation based tools lack user-flexibility as they are hardcoded to work only in certain genomes like humans or mouse primarily. This hampers researchers working with uncommon model organisms. Only very few tools, like sRNAtoolbox [28], Oasis [4], and ncPRO-Seq [7], do not have a hard-coded genome constraint, but they lack diverse graphical representation of data. In addition, existing tools are not tailored to compare multiple samples in a systematic way, properly normalizing sRNA datasets, thus allowing for an unbiased analysis. A non-exhaustive list of available sRNA analysis tools, and their abilities in addressing various properties essential to understand sRNA biogenesis, and mechanisms are discussed in Table 1. Inspite of the diverse availability of sRNA analysis tools, they have potential mishaps and do not capture all the qualitative, and quantitative properties (discussed in Table 1) while equipping the user with unbiased multi-sample comparisons.

Hence, we developed a generic sRNA analysis offline tool: Read Alignment, Analysis, and Differential PIpeline (RAPID), primarily tailored to investigate eukaryotic siRNAs. RAPID quantifies the basic alignment statistics with respect to read length, strand bias, non-templated nucleotides, nucleotide content, sequencing coverage etc. for user-defined sets of genes or regions of any reference genome. Once basic statistics are computed for multiple sRNA datasets, our tool aids the user with versatile functionalities, ranging from general quantitative analysis to visual comparison of multiple sRNA datasets.

# Materials and Methods

Figure 1 shows an overview of the various modules of RAPID, which we discuss below.

## Basic module

The first of four RAPID modules is *rapidStats*, which performs sequence (FastQ) alignment, with or without contaminant removal, using Bowtie2 [17]. After alignment, RAPID obtains read statistics such as read length distribution, soft-clipped nucleotides, strandedness, and nucleotide content. RAPID can skip the alignment and directly use alignment files (BAM/SAM) as well. To efficiently process, capture and store the aforementioned statistics, RAPID uses SAMtools [18], BEDtools [26], and custom Perl, Shell, and R scripts. The statistics captured by this module serve as input for other modules.

## Normalization module for multi-sample comparison

RAPID aims to facilitate an unbiased comparison of genes or regions across multiple sRNA samples. Other than the sequencing depth itself, sRNA studies pose an additional challenge during normalization. For instance, to understand RNA interference (RNAi) mechanisms and how the siRNA homeostasis is maintained, often a gene or siRNA region is knocked down. One such knockdown strategy is to introduce large amounts of siRNAs, called primary siRNAs, against the knockdown gene or any siRNA region. Consequentially, secondary siRNA production is triggered by the primary siRNAs. These primary and secondary siRNAs, which are also sequenced, can add up to millions of reads in the total library size.

To our knowledge, there are no normalization methods specialized for knockdown based small RNA-seq studies. However, many methods have been proposed to normalize mRNA-seq data, which can be broadly categorized in two classes: (i) total count scaling (TCS) methods and (ii) methods which utilize quantities like median log-fold change, among all genes between mRNA-seq experiments. To be able to use the latter methods, sRNA loci annotation should be available, and should assume that most of the sRNA loci between samples are not differentially expressed. In model organisms like *Paramecium tetraurelia*, little is known about the localization, and expression variability of endogenous small RNA loci. Hence, the second class of methods may not be applicable. However, the disadvantage of TCS methods is that the used normalization factors were shown to be biased by highly expressed genes in the dataset [8]. In case of knockdown samples, TCS methods will be heavily biased because of the millions of primary, and secondary siRNAs associated with the knockdown gene or region.

In mRNA-seq data, a variant of TCS method [30] was introduced, where normalization is achieved by scaling through a factor that estimates the difference in the number of reads mapped between samples. We previously proposed a variant of the TCS method in a knockdown based siRNA study [11]. Here, we term this variant as KnockDown Corrected Scaling (KDCS) method, where we remove from the estimated total library size, all small RNA reads that map against the knockdown genes, this quantity is denoted $K$ below. Assume read count $R$ for a region of interest that we want to compare between samples. $T$ is the total number of reads mapping to the genome, and $K$ is the number of small RNA reads mapping to the knockdown gene. We compute the normalized read count $\hat{R}$:

$$\hat{R} = R \cdot \frac{M}{T - K}, \tag{1}$$

where $M$ is the maximum over all values $(T_1 - K_1), ..., (T_n - K_n)$ over all $n$ samples. RAPID uses the KDCS method, by default. Hence, in the absence of knockdown genes, the normalization works as the normal TCS method. However, in order to provide flexibility with the choice of normalization for knockdown free analysis,we have also incorporated size factor-based normalization from DESeq2 [19]. If an user can safely assume that most of the genes or regions between samples are not differentially expressed, in a small RNA based study, then they can use the DESeq2 normalization.

## Visualization module

As visualization enables better understanding of data, the *rapidVis* module of RAPID automatically generates insightful plots from the output of previous modules. RAPID makes use of Rmarkdown (http://rmarkdown.rstudio.com) to create easily navigable HTML reports. This module contains two modes: *statistics* and *comparison* mode. The statistics mode accepts input from the *rapidStats* output file, and provides various single category plots detailing on the distribution of read length, strandedness, soft-clipped nucleotides, and coverage plots for each gene/region analyzed. In addition, this report also provides combinations of the aforementioned properties. For instance, how does strandedness differ across different read lengths. Comparison mode accepts the *rapidNorm* analysis output file, to equip the user with qualitative reports (Heatmaps, PCA, MDS) of samples. Further, sample and gene/region wise comparison plots of the properties inherent in the data. All plots are shown both in normal and log scale such that the user can directly incorporate them into publications.

## Differential analysis module

Differential Expression (DE) analysis is one of the common downstream analysis in comparative studies. RAPID equips the user with this functionality by incorporating the DESeq2 package. Upon invoking the *rapidDiff* module, raw counts are utilized from the output of the *rapidStats* module to perform DE analysis, with default parameters of DESeq2. Results of the DE analysis include intuitive plots (such as MA Plot, Heatmap, PCA) and the list of DE genes/regions.

## Usage and availability

We strongly recommend using RAPID from https://anaconda.org/bioconda/rapid as a conda recipe. However, it can also be freely accessed from https://github.com/SchulzLab/RAPID. A detailed use case based documentation is provided at http://rapid-doc.readthedocs.io/en/latest/.

## Definitions

Here we describe the formula used in the case studies. Coefficient of variation (CoV): Coefficient of variation is the ratio of standard deviation to mean of the data set. For a gene or region of interest, $i$, with $n$ samples

$$CoV_i = \frac{\sigma_i}{\mu_i}, \tag{2}$$

where $\sigma_i$, and $\mu_i$ are the standard deviation, and mean of the gene or region of interest $i$ in the $n$ samples respectively.

AntiSense Ratio (ASR):
Antisense ratio is the ratio of the number of antisense reads to the total number of reads in a gene or region. If $R$, and $AS$ are the total, and antisense read counts of a region of interest, $i$, respectively, antisense ratio is calculated as follows:

$$ASR_i = \frac{AS_i}{R_i}. \tag{3}$$

## DatasSets

We show the application of RAPID, using two different datasets which are briefly described below.

### *Paramecium tetraurelia*

We used four small RNA sequencing data sets (European Nucleotide Archive (ENA) Accession : PRJEB25903) from the wildtype serotypes (51A, 51B, 51D, and 51H) of *P.tetraurelia*. We performed adapter-trimming, merged the replicates, and extracted reads of length 21-25nt only from each dataset for this analysis. We analyzed only the rDNA cluster producing 17S, 5.8S, 25S ribosomal RNAs, External Transcribed Spacer (ETS), Internal Transcribed Spacer 1(ITS1), and Internal Transcribed Spacer 2 (ITS2). The rDNA cluster sequence can be obtained from GenBank Accession: AF149979.1 [24], with the additional annotation of the 5.8S sequence from GenBank accession: AM072801.1 [2].

In addition, to demonstrate a simple effect of KDCS normalization, we utilized the five available ICL knockdown data sets from this study (NCBI Accession ID: PRJEB13116) [11]. After preprocessing the data as mentioned in the study, we chose four small RNA regions as examples (They are regions of ND169 gene, as shown in the study) to quantify, and compare their sRNA accumulation across samples using RAPID.

### *Schizosaccharomyces pombe*

We explored the 24h time point datasets of WT, and three different knockdowns of *S.pombe* The respective data sets can be obtained by the accession ids: GEO:GSE89151; GSM2359756 - wt_24h; GSM2359762 - ago1D_24h; GSM2359768 - clr4D_24h; GSM2359774 - dcr1D_24h. We processed the data as mentioned in the corresponding download pages, before subjecting them to RAPID. We restricted our analysis to the sRNA-enriched genes available from their supplement [15] which can be accessed from `https://bit.ly/2GSLcks`. After pre-processing we subjected each GSM dataset to *rapidStats*, and compared them using *rapidNorm*.

# Results

We show the application of RAPID on two different datasets, highlighting some features that can be investigated by doing a standard RAPID analysis.

## Comparison of *Paramecium tetraurelia* serotypes

*P.tetraurelia* is a unicellular, free-living ciliate commonly found in fresh-water lakes. They show nuclear dimorphism, and have a wide range of phenotypes. These phenotypes depend on an epigenetically controlled, mutually exclusive expression of members of a multigene family called serotypes [6]. With a dimorphic nucleus and more than 11 serotypes, *P.tetraurelia* sRNAs are involved in regulatory mechanisms at the post-transcriptional, and epigenetic level [11, 6, 5].

Our first example is an analysis on four sRNA-seq datasets (ENA:PRJEB25903) from wildtype serotypes (51A, 51B, 51D, and 51H) of *P. tetraurelia*. We were interested in sRNAs produced in the rDNA cluster producing 17S, 5.8S, 25S ribosomal RNAs. A simple genomic visualization of the different components of rDNA cluster regions we quantified can be seen in Figure 2. We can observe from the strand-specific read distribution plots in Figure 2 that the regions, namely External Transcribed Spacer (ETS), Internal Transcribed Spacer 1(ITS1), and Internal Transcribed Spacer 2 (ITS2), which get excised in the processing of polycistronic pre-rRNA, accumulate 23nt antisense small RNAs. It is known from yeast, that rRNA maturation involves co-transcriptional endonucleolytic cleavage and highly concerted trimming events to subsequently process the final rRNAs [13]. Our data here suggests that in *P. tetraurelia* these elimination processes are associated with antisense siRNAs, possibly produced from RNA-dependent RNA Polymerase activity.

The commands, and visualization of the case studies, can be found in the readthedocs pages of RAPID at https://rapid-doc.readthedocs.io/en/latest/UseCases.html, and figures can be reproduced with the help of supplementary material.

## Normalization case study in *Paramecium tetraurelia* knockdowns

One of the unique features of RAPID is the KDCS normalization that can correct for the excess of sRNAs introduced in knockdown experiments in experimental approaches utilized in many diverse organisms. To demonstrate the effect of KDCS normalization, we utilized the ICL knockdown data sets from the study by [11]. This study investigates the molecular mechanisms of different sets of trans-acting RNAi components in *P. tetraurelia*. ICL is a gene in *P. tetraurelia*, which is not involved in the RNAi machinery. In the original study, as a control the ICL gene is knocked down by introducing primary siRNAs against ICL (see Methods). In our work, we quantified the sRNA read counts of four example sRNA regions (which are in the original study, different constructs of the ND169 gene) from the mentioned datasets. In this setup we expect that all datasets behave the same, as these are biological replicates of the same system.

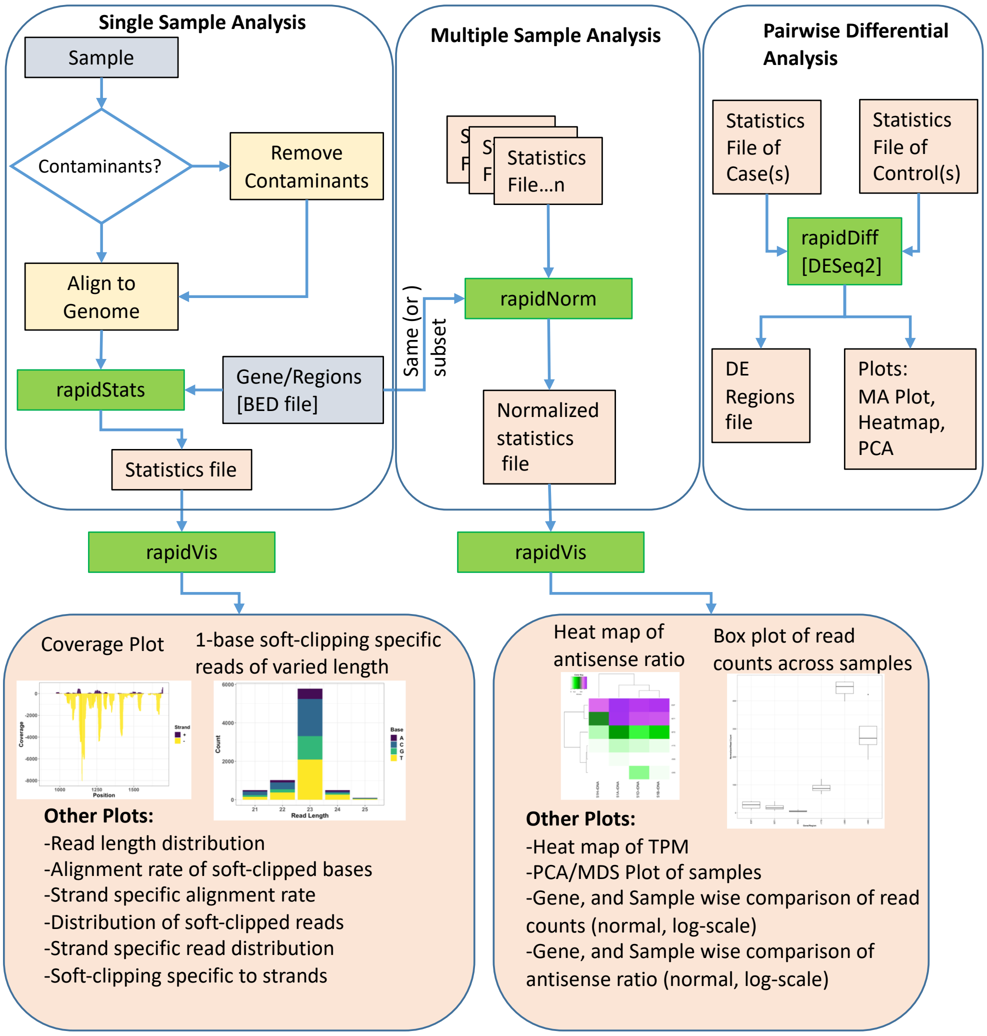
Single Sample Analysis
Sample
Contaminants?
Remove Contaminants
Align to Genome
rapidStats
Gene/Regions [BED file]
Statistics file
Multiple Sample Analysis
Statistics File…n
Same (or ) subset
rapidNorm
Normalized statistics file
Pairwise Differential Analysis
Statistics File of Case(s)
Statistics File of Control(s)
rapidDiff [DESeq2]
DE Regions file
Plots: MA Plot, Heatmap, PCA
rapidVis
rapidVis
Coverage Plot
1-base soft-clipping specific reads of varied length
Position
Coverage
Strand
Read Length
Count
Base
Other Plots:
-Read length distribution
-Alignment rate of soft-clipped bases
-Strand specific alignment rate
-Distribution of soft-clipped reads
-Strand specific read distribution
-Soft-clipping specific to strands
Heat map of antisense ratio
Box plot of read counts across samples
Other Plots:
-Heat map of TPM
-PCA/MDS Plot of samples
-Gene, and Sample wise comparison of read counts (normal, log-scale)
-Gene, and Sample wise comparison of antisense ratio (normal, log-scale)

Figure 1: The Pipeline of our tool RAPID is depicted. Green boxes are executables, Blue, and orange boxes represent input, and output files respectively. The executable RAPID modules are: (i) *rapidStats* module performs reference alignment and quantifies the expression of user-defined genes and/or regions. (ii) *rapidNorm* facilitates sample (or gene) wise comparison of genes/regions (or samples) after appropriate normalization. (iii) The *rapidVis* module provides multiple visualizations representing the information obtained from *rapidStats* and *rapidNorm*. Selective screenshots from the output of our case studies are shown in the boxes. (iv) *rapidDiff* is the differential expression analysis module implementing DESeq2.

As described earlier, very little is known about the localization, and expression variability of endogenous small RNA loci in *P. tetraurelia*. Therefore, in these sRNA knockdown samples, normalization methods, such as DESeq2, may be inappropriate, due to the assumption that the majority of regions are unchanged. We compared the effect of the TCS, DESeq2 and KDCS normalization approaches to using no normalization. We used the coefficient of variation (CoV) to measure the performance of the normalization method (see Methods; Equation 2) as normalization should reduce the variance in read count per region. A smaller CoV suggests a better performance of the normalization method.

Figure 3 shows the CoV values of the raw, and normalized sRNA read counts, for four example regions that had been studied by [11]. We can observe from Figure 3, that the KDCS method performs better in all the regions, compared to the generic TCS method. It also performs as good or better than the normalization of DESeq2 for this example. All normalization approaches are better than using no normalization, which strongly argues for their use. This experiment suggests that our KDCS method is a better alternative to the TCS method and is applicable when few regions are known.

## Analysis of *Schizosaccharomyces pombe* knockdowns

The fission yeast, or *Schizosaccharomyces pombe*, is another widely studied unicellular eukaryotic model organism, where RNAi pathways are prevalent. We explored the time point datasets of WT, and three different knockdowns of *S.pombe* from the study by [15]. In this study, the authors investigated quiescence associated changes in small RNA transcriptomes and epigenetic modifications to identify the key players involved in quiescence. They also examined the role of RNAi proteins, by knocking down three of them, namely, Ago1, Clr4, and Dcr1. One of the key findings was that during quiescence in *S.pombe*, a set of sRNA-enriched genes were identified as crucial elements for the survival of the organism [15]. We explored these sRNA-enriched genes using RAPID, to demonstrate the discovering potential of RAPID.

With a simple RAPID analysis comparing the different knockdown samples, it was easy to screen for interesting properties in the data. We discovered that a subset of these sRNA-enriched genes have relatively higher antisense ratio (ASR, see Methods Equation 3) (Figure 4). This increased ASR was observed in different sets of genes in various knockdowns. The subgroup of genes with higher ASR, could play a *cis*-regulatory role to silence the genes. While further investigation is necessary, our RAPID analysis (Figure 4) suggests an involvement of different sRNA mechanisms in ensuring the survival of *S.pombe* in quiescence.

# Discussion

The long list of available sRNA analysis tools attributes to the complexity, and importance of sRNAs in biological studies. However, most available tools only focus on identifying, and annotating the different classes of sRNA. They fail to characterize and visually represent the multitude of parameters crucial for understanding the sRNA world.

RAPID is designed to capture the diverse eukaryotic siRNA characteristics innately found in sRNA sequencing data sets. Some of the properties captured during the basic analysis of RAPID include read length, strand bias, non-templated nucleotides, nucleotide content, sequencing coverage etc. for user-defined sets of genes (or regions) of any reference genome. With a separate module for normalization, RAPID simplifies multi-sample comparison. We have also included an alternative normalization technique, KDCS, specially designed to aid the comparison of sRNA-based knockdown studies. KDCS normalization method can also be helpful in correcting for the transcribed small RNAs from the non-insert RNA locations of a vector. For instance, in RNAi vector constructs, like L4440 in *Caenorhabditis elegans*, due to lack of specificity of the termination enzyme, the non-insert RNA locations will get

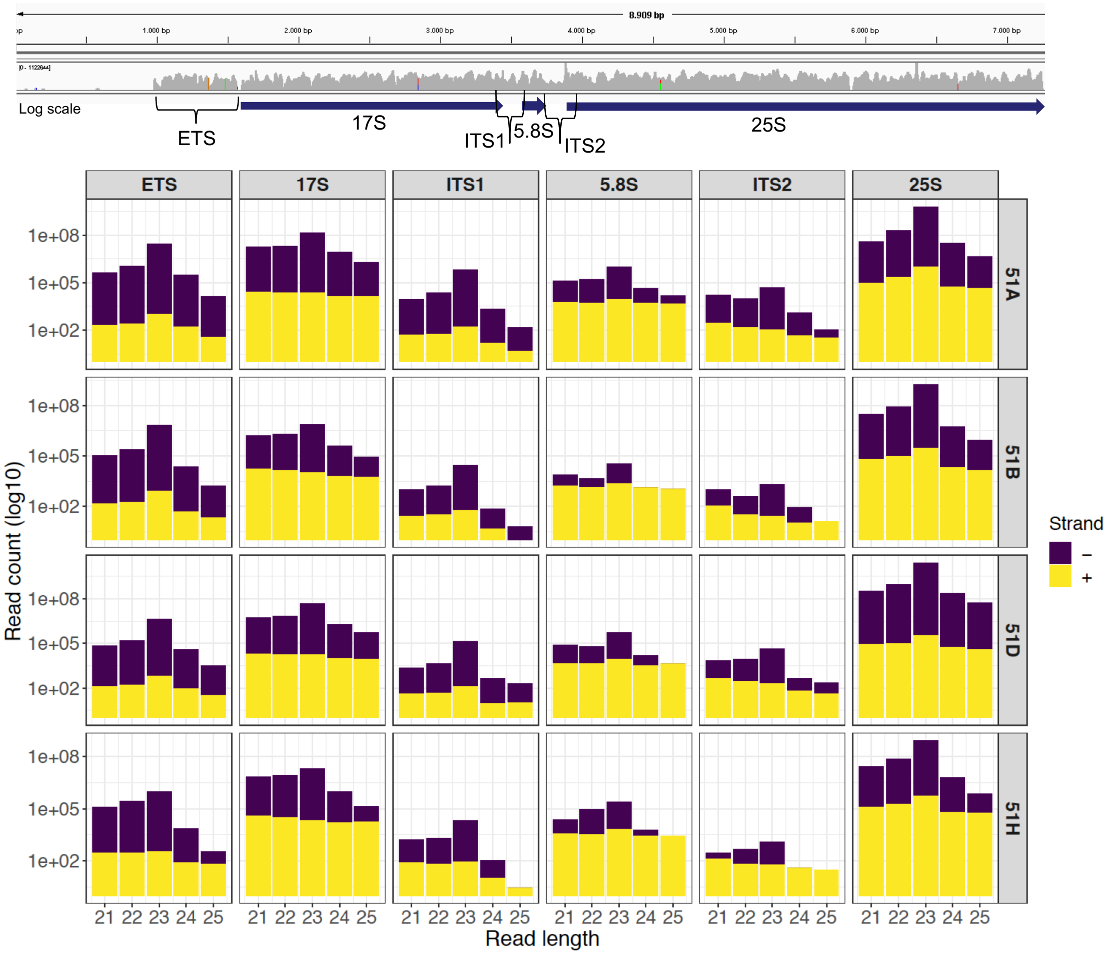

Figure 2: The read length distribution of small RNAs mapping to different regions of the ribosomal DNA is shown. Top: IGV screenshots of genomic localization of ribosomal DNA. Bottom: Each row corresponds to each wildtype serotype (51A, 51B, 51D, 51H respectively). Each column corresponds to ribosomal DNA regions in order: External Transcribed Spacer (ETS), 17S, Internal Transcribed Spacer 1(ITS1), 5.8S, Internal Transcribed Spacer 2 (ITS2), and 25S

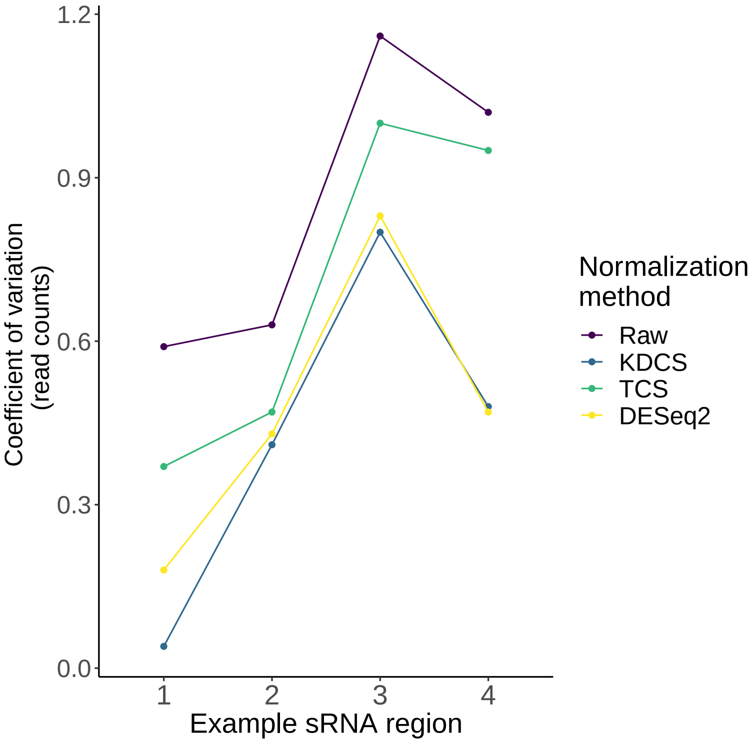

Figure 3: The effect of different normalization methods on sRNA regions (x-axis) studied in [11] are assessed using the coefficient of variation (y-axis; lower is better) of the read counts obtained from RAPID. Raw - No normalization; KDCS - KnockDown Corrected Scaling; TCS - Total Count Scaling; DESeq2 - size factor-based normalization from DESeq2

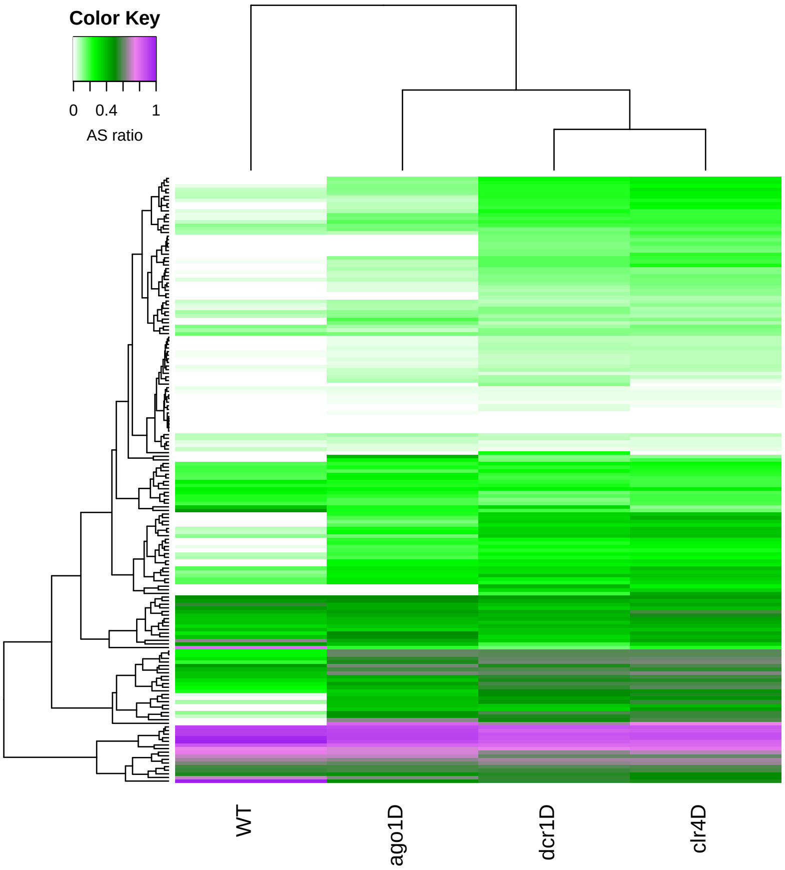

Figure 4: A heatmap of the antisense read ratio in the sRNA enriched genes (y-axis) across all samples (x-axis) analyzed. Accumulation of antisense sRNA can be observed in the lower part of the heatmap, and an increase in antisense sRNA can also be seen in different knockdowns compared to wildtype.

transcribed [29]. These regions which contribute unwanted variation can be excluded by specifying them as background in the RAPID analysis.

RAPID currently addresses many features which are crucial towards the understanding of sRNA biogenesis, and function. In spite of RAPID's diverse functionality, there are a few shortcomings. RAPID depends on user supplied set of contaminants instead of auto-detecting it from the sequence file. The visualizations provided by RAPID in the statistics mode do not include sequence level properties, like over represented sequences or sequence logos, etc. These are interesting additions for future releases. In the comparison mode of our visualization module, the plots provided are non-exhaustive. For instance, one might like to compare the nucleotide content, or strand-based distribution of genes (or regions) across multiple samples, or vice-versa. Such special features can be requested by users in GitHub, which could be incorporated in further releases.

# Conclusion

RAPID is an offline, open-source, user-friendly, and automated pipeline designed to simplify sRNA data analysis. RAPID is not an exhaustive sRNA analysis or annotation pipeline. With an available set of sRNA localizations, our tool can be used to analyze single or multiple sRNA samples at ease with the aid of different normalization techniques. The diverse set of visualizations generated by RAPID will enhance the understanding of any sRNA-based study. RAPID is available for free use and can be used over the command line. It is available at the github repository (https://github.com/SchulzLab/RAPID). A detailed user-tutorial can be accessed from this repository.

# Acknowledgements

We would like to thank Miriam Cheaib, Angela Rodrigues, and Fatemeh Behjati Ardakani for their valuable feedback.

# Funding

This work was supported by grants from the German research council (DFG) to MS (SI1379/3-1) and to MHS SCHU (3140/1-1).

# References

[1] M. J. Axtell. ShortStack: Comprehensive annotation and quantification of small RNA genes. *RNA*, 19:740–751, 2013.

[2] D. Barth, S. Krenek, S. I. Fokin, and T. U. Berendonk. Intraspecific genetic variation in Paramecium revealed by mitochondrial cytochrome c oxidase I sequences. *Journal of Eukaryotic Microbiology*, 53(1):20–25, 2006.

[3] Lionello Bossi and Nara Figueroa-Bossi. Competing endogenous RNAs: A target-centric view of small RNA regulation in bacteria. *Nature Reviews Microbiology*, 14:775–784, 2016.

[4] V. Capece, J. C. Garcia Vizcaino, R. Vidal, R.-U. Rahman, T. Pena Centeno, O. Shomroni, I. Suberviola, a. Fischer, and S. Bonn. Oasis: online analysis of small RNA deep sequencing data. *Bioinformatics*, 31(13):2205–2207, 2015.

[5] Q. Carradec, U. Götz, O. Arnaiz, J. Pouch, M. Simon, E. Meyer, and S. Marker. Primary and secondary siRNA synthesis triggered by RNAs from food bacteria in the ciliate Paramecium tetraurelia. *Nucleic Acids Research*, 43(3):1818–33, 2015.

| | | | | | | | Qualitative Plots | | | Quantitative Plots | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tool/Supporting Feature | Contaminant removal | Supports other aligners | User-defined gene/region | Knockdown corrected normalization | Offline | Hardcoded genomes | Heatmaps | PCA | MDS | Strand specificity | Soft-clipped bases | Coverage plots | Multi-sample comparison plots | Differential analysis | Enrichment analysis support | Interactive interface | miRNA or piRNA specific? |
| RAPID | ✓ | ✓ | ✓ | ✓ | ✓ | x | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | x | x | x |
| smallRNA toolkit[21] | x | x | ✓ | NA | x | ✓ | x | x | x | x | x | x | x | x | x | x | ✓ |
| sRNA toolbox[28] | x | x | ✓ | NA | ✓ | x | x | x | x | x | x | x | x | ✓ | ✓ | x | ✓ |
| Oasis[4] | x | x | x | x | x | x | ✓ | ✓ | x | x | x | x | ✓ | ✓ | ✓ | ✓ | ✓ |
| CPSS[31] | x | ✓ | x | NA | x | ✓ | x | x | x | x | x | x | x | x | ✓ | ✓ | ✓ |
| iSmart[22] | x | x | x | NA | ✓ | ✓ | ✓ | ✓ | x | x | x | x | x | ✓ | ✓ | ✓ | ✓ |
| iSRAP[25] | x | ✓ | x | NA | ✓ | x | ✓ | ✓ | ✓ | x | x | x | x | ✓ | x | x | x |
| PiPipes[12] | ✓ | ✓ | x | NA | ✓ | x | x | x | x | ✓ | ✓ | x | x | ✓ | x | x | ✓ |
| ncPRO-Seq[7] | x | ✓ | ✓ | NA | ✓ | x | x | x | x | x | x | x | x | x | ✓ | x | ✓ |
| UEA sRNAworkbench [20] | ✓ | ✓ | ✓ | x | ✓ | x | x | x | x | x | x | x | ✓ | ✓ | x | ✓ | x |
| NGSToolbox [27] | x | x | ✓ | NA | ✓ | x | x | x | x | x | x | x | x | x | x | x | ✓ |
| SePIA [14] | x | ✓ | x | NA | ✓ | x | x | x | x | x | x | x | x | ✓ | x | x | ✓ |
| SPAR [16] | x | ✓ | x | NA | x | ✓ | ✓ | x | x | x | x | x | x | x | x | ✓ | ✓ |

Table 1: Comparison of RAPID with other tools is shown. ✓- Feature supported, x - Feature not supported, NA - Feature is not in the scope of this tool. For instance, *Knockdown corrected normalization* feature is *NA* for CPSS, because it does not support multiple sample comparison. The full description of the column headers are listed in Table 2. *Note*: Tools whose primary focus is on identifying/annotating different classes of small RNAs are not included.

| **Supporting Feature** | **Description** |
|---|---|
| Contaminant removal | Is there an option to remove set of contaminants (microbial, ribosomal, etc.) from the read files? |
| Supports other aligners | Does the tool support alignment files from other tools, instead of performing their own alignment? |
| User-defined gene/region | Could the user specify a list of regions to perform downstream analysis? |
| Knockdown corrected normalization | Does the tool enable multiple-sample comparison by facilitating normalization techniques specific to sRNA knockdown studies? |
| Offline | Can the tool be used offline? |
| Hardcoded genomes | Is the tool generic? i.e. Is the tool's ability somehow limited to a set of pre-defined genomes? |
| Quantitative, and Qualitative Plots | Does the tool support informative plots to gain understanding of the analyzed data (MDS=multi-dimensional scaling, PCA= principal component analysis) |
| Multi-sample comparison plots | Does the tool provide a comprehensive view of multiple samples (not just differential analysis)? For instance, how does the read distribution vary across multiple samples in different genes of interest? |
| Differential analysis | Is the tool equipped with modules to perform pairwise differential analysis? |
| Enrichment analysis support | Is there any support to perform functional enrichment within the tool |
| Interactive interface | Does the tool have an interactive interface, or plots? |
| miRNA or piRNA specific? | Is the tool specific to analyze miRNA or piRNA only? |

Table 2: Table describing the supporting features of RAPID (Column headers in Table 1).

[6] M. Cheaib, A. Dehghani Amirabad, K. J. Nordström, M. H. Schulz, and M. Simon. Epigenetic regulation of serotype expression antagonizes transcriptome dynamics in Paramecium tetraurelia. *DNA Research*, 22(4):293–305, 2015.

[7] Chong Jian Chen, Nicolas Servant, Joern Toedling, Alexis Sarazin, Antonin Marchais, Evelyne Duvernois-Berthet, Valérie Cognat, Vincent Colot, Olivier Voinnet, Edith Heard, Constance Ciaudo, and Emmanuel Barillot. NcPRO-seq: A tool for annotation and profiling of ncRNAs in sRNA-seq data. *Bioinformatics*, 28(23):3147–3149, 2012.

[8] Marie-Agnès Dillies, , Andrea Rau, , Julie Aubert, , Christelle Hennequet-Antier, , Marine Jeanmougin, , Nicolas Servant, , Céline Keime, , Guillemette Marot, , David Castel, , Jordi Estelle, , Gregory Guernec, , Bernd Jagla, , Luc Jouneau, , Denis Laloë, , Caroline Le Gall, , Brigitte Schaëffer, , Stéphane Le Crom, , Mickaël Guedj, , Florence Jaffrézic, and . A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, 2013.

[9] Marc R. Friedländer, Sebastian D. MacKowiak, Na Li, Wei Chen, and Nikolaus Rajewsky. MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, 40(1):37–52, 2012.

[10] Giorgio Giurato, Maria R. De Filippo, Antonio Rinaldi, Adnan Hashim, Giovanni Nassa, Maria Ravo, Francesca Rizzo, Roberta Tarallo, and Alessandro Weisz. IMir: An integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq. *BMC Bioinformatics*, 14:362, 2013.

[11] Ulrike Götz, Simone Marker, Miriam Cheaib, Karsten Andresen, Simon Shrestha, Dilip A. Durai, Karl J. Nordstrom, Marcel H. Schulz, and Martin Simon. Two sets of RNAi components are required for heterochromatin formation in trans triggered by truncated transgenes. *Nucleic Acids Research*, 44:5908–5923, 2016.

[12] Bo W. Han, Wei Wang, Phillip D. Zamore, and Zhiping Weng. PiPipes: A set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. *Bioinformatics*, 31(4):593–595, 2015.

[13] Anthony K. Henras, Célia Plisson-Chastang, Marie Françoise O'Donohue, Anirban Chakraborty, and Pierre Emmanuel Gleizes. An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdisciplinary Reviews: RNA*, 6:225–242, 2015.

[14] Katherine Icay, Ping Chen, Alejandra Cervera, Ville Rantanen, Rainer Lehtonen, and Sampsa Hautaniemi. SePIA: RNA and small RNA sequence processing, integration, and analysis. *BioData Mining*, 9:20, 2016.

[15] Richard I. Joh, Jasbeer S. Khanduja, Isabel A. Calvo, Meeta Mistry, Christina M. Palmieri, Andrej J. Savol, Shannan J. Ho Sui, Ruslan I. Sadreyev, Martin J. Aryee, and Mo Motamedi. Survival in Quiescence Requires the Euchromatic Deployment of Clr4/SUV39H by Argonaute-Associated Small RNAs. *Molecular Cell*, 64:1088–1101, 2016.

[16] Pavel P. Kuksa, Alexandre Amlie-Wolf, źivadin Katanić, Otto Valladares, Li San Wang, and Yuk Yee Leung. SPAR: Small RNA-seq portal for analysis of sequencing experiments. *Nucleic Acids Research*, 46(W1):W36–W42, 2018.

[17] Ben Langmead, Steven L Salzberg, and Langmead. Bowtie2. *Nature methods*, 9:357–359, 2013.

[18] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, Genome Project Data, The Sam, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment / Map format and SAMtools. *Bioinformatics*, 25:2078–2079, 2009.

[19] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, Dec 2014.

[20] Irina Mohorianu, Matthew Benedict Stocks, Christopher Steven Applegate, Leighton Folkes, and Vincent Moulton. The UEA small RNA workbench: A suite of computational tools for small RNA analysis. In *Methods in Molecular Biology*, volume 1580, pages 193–224. Springer, 2017.

[21] Simon Moxon, Frank Schwach, Tamas Dalmay, Dan MacLean, David J. Studholme, and Vincent Moulton. A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics*, 24(9):2252–2253, 2008.

[22] Riccardo Panero, Antonio Rinaldi, Domenico Memoli, Giovanni Nassa, Maria Ravo, Francesca Rizzo, Roberta Tarallo, Luciano Milanesi, Alessandro Weisz, and Giorgio Giurato. ISmaRT: A toolkit for a comprehensive analysis of small RNA-Seq data. *Bioinformatics*, 33(16):938–940, 2017.

[23] Bhushan Patwardhan, Rathnam Chaguturu, Preeti Chavan-Gautam, Tejas Shah, and Kalpana Joshi. Chapter 8 - Transcriptomics and Epigenomics. In *Innovative Approaches in Drug Discovery*, page 235. Elsevier, 2017.

[24] L B Preer, B Rudman, S Pollack, and J R Preer. Does ribosomal DNA get out of the micronuclear chromosome in Paramecium tetraurelia by means of a rolling circle? *Molecular and cellular biology*, 19(11):7792–7800, 1999.

[25] Camelia Quek, Chol hee Jung, Shayne A. Bellingham, Andrew Lonie, and Andrew F. Hill. iSRAP - A one-touch research tool for rapid profiling of small RNA-seq data. *Journal of Extracellular Vesicles*, 4:29454, 2015.

[26] Aaron R. Quinlan and Ira M. Hall. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26:841–842, 2010.

[27] David Rosenkranz, Chung Ting Han, Elke F. Roovers, Hans Zischler, and René F. Ketting. Piwi proteins and piRNAs in mammalian oocytes and early embryos: From sample to sequence. *Genomics Data*, 5:309–313, 2015.

[28] Antonio Rueda, Guillermo Barturen, Ricardo Lebrón, Cristina Gómez-Martín, Ángel Alganza, José L. Oliver, and Michael Hackenberg. SRNAtoolbox: An integrated collection of small RNA research tools. *Nucleic Acids Research*, 43(W1):W467–W473, 2015.

[29] Éva Saskói, Kovács Tibor, Ádám Sturm, Tibor Vellai, and Nóra Weinhardt. Highly efficient RNAi and Cas9-based auto-cloning systems for C. elegans research. *Nucleic Acids Research*, 46(17):e105–e105, 06 2018.

[30] Marc Sultan, Marcel H. Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, Tatjana Borodina, Aleksey Soldatov, Dmitri Parkhomchuk, Dominic Schmidt, Sean O'Keeffe, Stefan Haas, Martin Vingron, Hans Lehrach, and Marie-Laure Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960, 2008.

[31] Changlin Wan, Jianing Gao, Huan Zhang, Xiaohua Jiang, Qiguang Zang, Rongjun Ban, Yuanwei Zhang, and Qinghua Shi. CPSS 2.0: A computational platform update for the analysis of small RNA sequencing data. *Bioinformatics*, 33(20):3289–3291, 2017.

[32] Kai Wang, Chun Liang, Jinding Liu, Huamei Xiao, Shuiqing Huang, Jianhua Xu, and Fei Li. Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinformatics*, 15:419, 2014.

[33] Jinyu Wu, Qi Liu, Xin Wang, Jiayong Zheng, Tao Wang, Mingcong You, Zhong Sheng Sun, and Qinghua Shi. MirTools 2.0 for non-coding RNA discovery, profiling and functional annotation based on high-throughput sequencing. *RNA Biology*, 10(7):1087–1092, 2013.