Automated analysis of small RNA datasets with RAPID

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

[1]**Cluster of Excellence for Multimodal Computing and Interaction, and Department for Computational Biology & Applied Algorithmics, Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany.**

[2]**Graduate School of Computer Science, Saarland Informatics Campus, Universität des Saarlandes, Saarbrücken, Germany.**

[3]**Institute for Cardiovascular Regeneration, Goethe University Hospital, Frankfurt am Main, Germany**

[4]**Molecular Cell Biology and Microbiology, Wuppertal University, Wuppertal, Germany**

[5]**German Centre for Cardiovascular Research (DZHK), Partner site RheinMain, Frankfurt am Main, Germany**

Corresponding author:

Marcel H. Schulz[1]

Email address: marcel.schulz@em.uni-frankfurt.de

ABSTRACT

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.

Availability and Implementation: 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 (Bossi and Figueroa-Bossi, 2016). 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 (Patwardhan et al., 2017). 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

47 includes methods such as Shortstack (Axtell, 2013), miRDeep2 (Friedländer et al., 2012), iMir (Giurato
48 et al., 2013), Piano (Wang et al., 2014), etc. (ii) The secondary focus of many sRNA analysis tools
49 is to annotate, and perform Gene Ontology (GO) enrichment analysis of known, or predicted sRNAs.
50 Examples include miRTools2 (Wu et al., 2013), iSmart (Panero et al., 2017), and CPSS (Wan et al., 2017).
51 However, such annotation based tools lack user-flexibility as they are hardcoded to work only in certain
52 genomes like humans or mouse primarily. This hampers researchers working with uncommon model
53 organisms. Only very few tools, like sRNAtoolbox (Rueda et al., 2015), Oasis (Capece et al., 2015),
54 and ncPRO-Seq (Chen et al., 2012), do not have a hard-coded genome constraint, but they lack diverse
55 graphical representation of data. In addition, existing tools are not tailored to compare multiple samples
56 in a systematic way, properly normalizing sRNA datasets, thus allowing for an unbiased analysis. A
57 non-exhaustive list of available sRNA analysis tools, and their abilities in addressing various properties
58 essential to understand sRNA biogenesis, and mechanisms are discussed in Table 1. Inspite of the diverse
59 availability of sRNA analysis tools, they have potential mishaps and do not capture all the qualitative,
60 and quantitative properties (discussed in Table 1, and Table 2) while equipping the user with unbiased
61 multi-sample comparisons.

62

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

MATERIALS AND METHODS

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

Basic module

72 The first of four RAPID modules is *rapidStats*, which performs sequence (FastQ) alignment, with or
73 without contaminant removal, using Bowtie2 (Langmead et al., 2013). After alignment, RAPID obtains
74 read statistics such as read length distribution, soft-clipped nucleotides, strandedness, and nucleotide
75 content. RAPID can skip the alignment and directly use alignment files (BAM/SAM) as well. To
76 efficiently process, capture and store the aforementioned statistics, RAPID uses SAMtools (Li et al.,
77 2009), BEDtools (Quinlan and Hall, 2010), and custom Perl, Shell, and R scripts. The statistics captured
78 by this module serve as input for other modules.

Normalization module for multi-sample comparison

80 RAPID aims to facilitate an unbiased comparison of genes or regions across multiple sRNA samples.
81 Other than the sequencing depth itself, sRNA studies pose an additional challenge during normalization.
82 For instance, to understand RNA interference (RNAi) mechanisms and how the siRNA homeostasis is
83 maintained, often a gene or siRNA region is knocked down. One such knockdown strategy is to introduce
84 large amounts of siRNAs, called primary siRNAs, against the knockdown gene or any siRNA region.
85 Consequentially, secondary siRNA production is triggered by the primary siRNAs. These primary and
86 secondary siRNAs, which are also sequenced, can add up to millions of reads in the total library size.
87 To our knowledge, there are no normalization methods specialized for knockdown based small RNA-
88 seq studies. However, many methods have been proposed to normalize mRNA-seq data, which can be
89 broadly categorized in two classes: (i) total count scaling (TCS) methods and (ii) methods which utilize
90 quantities like median log-fold change, among all genes between mRNA-seq experiments. To be able
91 to use the latter methods, sRNA loci annotation should be available, and should assume that most of
92 the sRNA loci between samples are not differentially expressed. In model organisms like *Paramecium*
93 *tetraurelia*, little is known about the localization, and expression variability of endogenous small RNA
94 loci. Hence, the second class of methods may not be applicable. However, the disadvantage of TCS
95 methods is that the used normalization factors were shown to be biased by highly expressed genes in the
96 dataset (Dillies et al., 2013). In case of knockdown samples, TCS methods will be heavily biased because
97 of the millions of primary, and secondary siRNAs associated with the knockdown gene or region.

2/13

In mRNA-seq data, a variant of TCS method (Sultan et al., 2008) 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 (Götz et al., 2016). 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 $\hat{R}$:

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

98 where M is the maximum over all values $(T_1 - K_1), ..., (T_n - K_n)$ over all n samples.

99 RAPID uses the KDCS method, by default. Hence, in the absence of knockdown genes, the normal-
100 ization works as the normal TCS method. However, in order to provide flexibility with the choice of
101 normalization for knockdown free analysis,we have also incorporated size factor-based normalization
102 from DESeq2 (Love et al., 2014). If an user can safely assume that most of the genes or regions between
103 samples are not differentially expressed, in a small RNA based study, then they can use the DESeq2
104 normalization.

Visualization module

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

Differential analysis module

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

Usage and availability

124 We strongly recommend using RAPID from https://anaconda.org/bioconda/rapid as a conda recipe.
125 However, it can also be freely accessed from https://github.com/SchulzLab/RAPID. A detailed use case
126 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 σ_i, and μ_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:

$$ASR_i = \frac{AS_i}{R_i}.$$ (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 : PR-JEB25903) 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 (Preer et al., 1999), with the additional annotation of the 5.8S sequence from GenBank accession: AM072801.1 (Barth et al., 2006).

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) (Götz et al., 2016). 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 (Joh et al., 2016) 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 (Cheaib et al., 2015). With a dimorphic nucleus and more than 11 serotypes, *P.tetraurelia* sRNAs are involved in regulatory mechanisms at the post-transcriptional, and epigenetic level (Götz et al., 2016; Cheaib et al., 2015; Carradec et al., 2015).

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 (Henras et al., 2015). Our data here suggests that in *P. tetraurelia* these elimination processes are associated with antisense siRNAs, possibly produced from RNA-dependent RNA Polymerase activity.
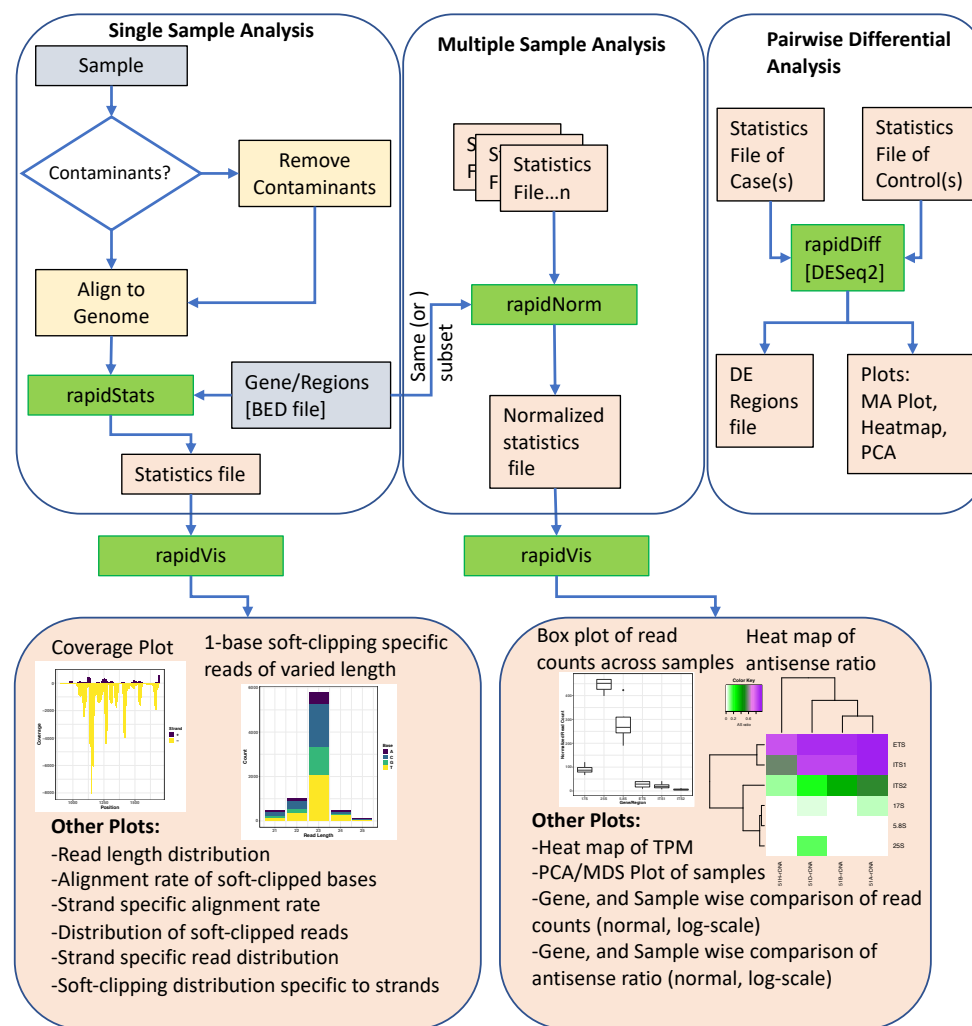
Figure 1. The Pipeline of our tool RAPID is depicted. Green boxes are executables, Blue, and orange boxes represent input, and output files respectively. The executable RAPID modules are: (i) *rapidStats* module performs reference alignment and quantifies the expression of user-defined genes and/or regions. (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.

173   Analysis of our case studies, and their respective figures can be reproduced with the help of supple-
174 mentary material (Data S1.zip).

Normalization case study in *Paramecium tetraurelia* knockdowns

176 One of the unique features of RAPID is the KDCS normalization that can correct for the excess of sRNAs
177 introduced in knockdown experiments in experimental approaches utilized in many diverse organisms. To
178 demonstrate the effect of KDCS normalization, we utilized the ICL knockdown data sets from the study
179 by Götz et al. (2016). This study investigates the molecular mechanisms of different sets of trans-acting
180 RNAi components in *P. tetraurelia*. ICL is a gene in *P. tetraurelia*, which is not involved in the RNAi
181 machinery. In the original study, as a control the ICL gene is knocked down by introducing primary
182 siRNAs against ICL (see Methods). In our work, we quantified the sRNA read counts of four example
183 sRNA regions (which are in the original study, different constructs of the ND169 gene) from the mentioned
184 datasets. In this setup we expect that all datasets behave the same, as these are biological replicates of the
185 same system.

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

193   Figure 3 shows the CoV values of the raw, and normalized sRNA read counts, for four example
194 regions that had been studied by Götz et al. (2016). We can observe from Figure 3, that the KDCS method
195 performs better in all the regions, compared to the generic TCS method. It also performs as good or better
196 than the normalization of DESeq2 for this example. All normalization approaches are better than using no
197 normalization, which strongly argues for their use. This experiment suggests that our KDCS method is a
198 better alternative to the TCS method and is applicable when few regions are known.

Analysis of *Schizosaccharomyces pombe* knockdowns

200 The fission yeast, or *Schizosaccharomyces pombe*, is another widely studied unicellular eukaryotic
201 model organism, where RNAi pathways are prevalent. We explored the time point datasets of WT, and
202 three different knockdowns of *S.pombe* from the study by (Joh et al., 2016). In this study, the authors
203 investigated quiescence associated changes in small RNA transcriptomes and epigenetic modifications
204 to identify the key players involved in quiescence. They also examined the role of RNAi proteins, by
205 knocking down three of them, namely, Ago1, Clr4, and Dcr1. One of the key findings was that during
206 quiescence in *S.pombe*, a set of sRNA-enriched genes were identified as crucial elements for the survival
207 of the organism (Joh et al., 2016). We explored these sRNA-enriched genes using RAPID, to demonstrate
208 the discovering potential of RAPID.

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

DISCUSSION

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

221   RAPID is designed to capture the diverse eukaryotic siRNA characteristics innately found in sRNA
222 sequencing data sets. Some of the properties captured during the basic analysis of RAPID include
223 read length, strand bias, non-templated nucleotides, nucleotide content, sequencing coverage etc. for
224 user-defined sets of genes (or regions) of any reference genome. With a separate module for normalization,
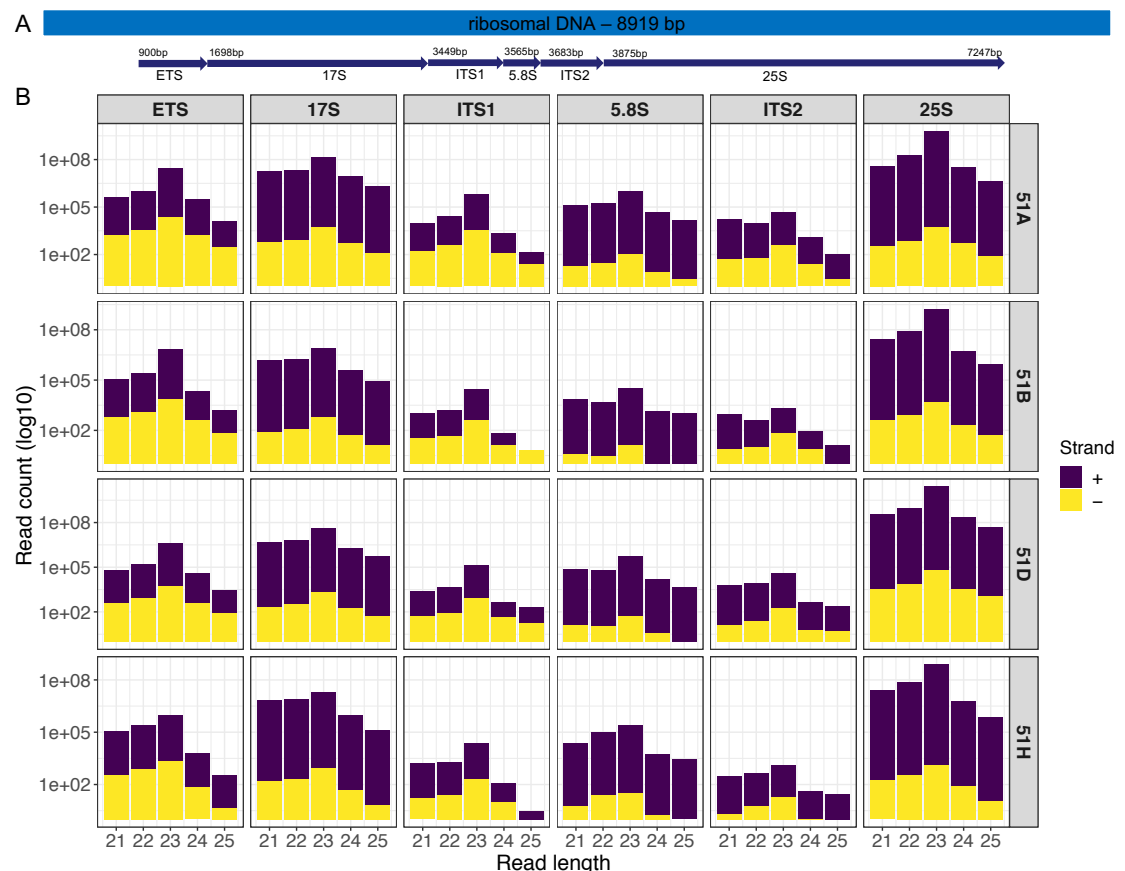
Figure 2. Small RNAs map to different units of the ribosomal DNA of *P.tetraurelia*. A) Genomic localization of different ribosomal units (Block arrows, with respective start positions) is shown. B) The read length distribution of small RNAs mapping to different regions of the ribosomal DNA is shown. 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
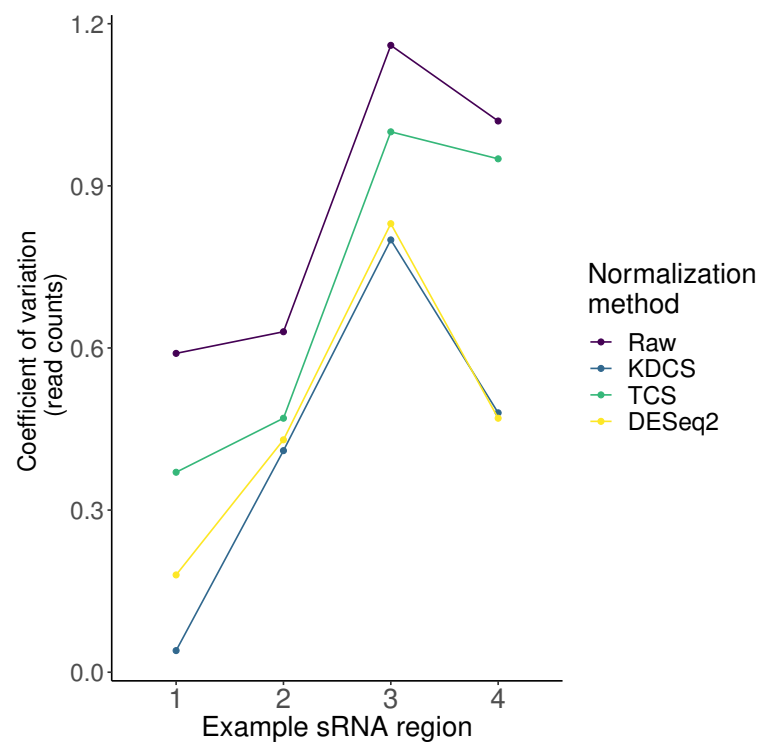
Figure 3. The effect of different normalization methods on sRNA regions (x-axis) studied in (Götz et al., 2016) are assessed using the coefficient of variation (y-axis; lower is better) of the read counts obtained from RAPID. Raw - No normalization; KDCS - KnockDown Corrected Scaling; TCS - Total Count Scaling; DESeq2 - size factor-based normalization from DESeq2
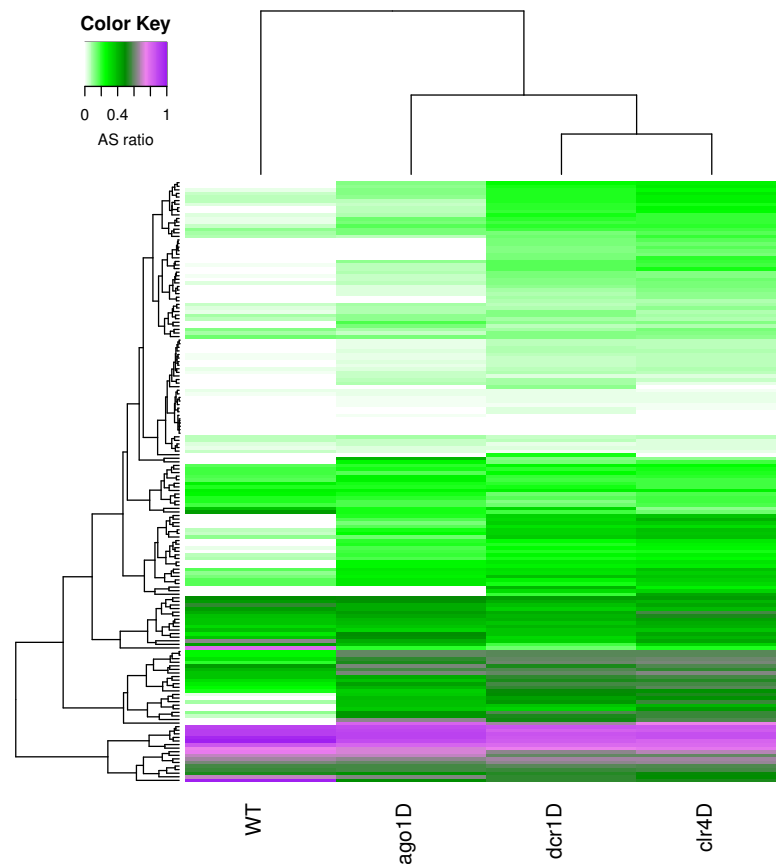
Figure 4. A heatmap of the antisense read ratio in the sRNA enriched genes (y-axis) across all samples (x-axis) analyzed. Accumulation of antisense sRNA can be observed in the lower part of the heatmap, and an increase in antisense sRNA can also be seen in different knockdowns compared to wildtype.

Tool/Supporting Feature	Contaminant removal	Supports other aligners	User-defined gene/region	Knockdown corrected normalization	Offline	Hardcoded genomes	Qualitative Plots Heatmaps	PCA	MDS	Quantitative Plots 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(Moxon et al., 2008)	x	x	✓	NA	x	✓	x	x	x	x	x	x	x	x	x	x	✓
sRNA toolbox(Rueda et al., 2015)	x	x	✓	NA	✓	x	x	x	x	x	x	x	x	✓	✓	x	✓
Oasis(Capece et al., 2015)	x	x	x	x	x	x	✓	✓	x	x	x	x	✓	✓	✓	✓	✓
CPSS(Wan et al., 2017)	x	✓	x	NA	✓	x	x	x	x	x	x	x	x	x	✓	✓	✓
iSmart(Panero et al., 2017)	x	x	x	NA	✓	✓	✓	✓	x	x	x	x	x	✓	✓	✓	✓
iSRAP(Quek et al., 2015)	x	✓	x	NA	✓	x	✓	✓	✓	x	x	x	x	✓	x	x	x
PiPipes(Han et al., 2015)	✓	✓	x	NA	✓	x	x	x	x	✓	✓	x	x	✓	x	x	✓
ncPRO-Seq(Chen et al., 2012)	x	✓	x	NA	✓	x	x	x	x	x	x	x	x	x	✓	x	✓
UEA sRNAworkbench (Mohorianu et al., 2017)	✓	✓	✓	x	✓	x	x	x	x	x	x	x	✓	x	✓	✓	x
NGSToolbox (Rosenkranz et al., 2015)	x	x	✓	NA	✓	x	x	x	x	x	x	x	x	x	x	x	✓
SePIA (Icay et al., 2016)	x	✓	x	NA	✓	x	x	x	x	x	x	x	x	✓	x	x	✓
SPAR (Kuksa et al., 2018)	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.

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 transcribed (Saskói et al., 2018). 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 siRNA 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 data analysis, tailored to investigate eukaryotic siRNAs. 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.

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).

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 Martin Simon (SI1379/3-1) and to Marcel H. Schulz (SCHU3140/1-1).

COMPETING INTERESTS

The authors declare there are no competing interests.

REFERENCES

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

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

Bossi, L. and Figueroa-Bossi, N. (2016). Competing endogenous RNAs: A target-centric view of small RNA regulation in bacteria. *Nature Reviews Microbiology*, 14:775–784.

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

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

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

Chen, C. J., Servant, N., Toedling, J., Sarazin, A., Marchais, A., Duvernois-Berthet, E., Cognat, V., Colot, V., Voinnet, O., Heard, E., Ciaudo, C., and Barillot, E. (2012). NcPRO-seq: A tool for annotation and profiling of ncRNAs in sRNA-seq data. *Bioinformatics*, 28(23):3147–3149.

Dillies, M.-A., , Rau, A., , Aubert, J., , Hennequet-Antier, C., , Jeanmougin, M., , Servant, N., , Keime, C., , Marot, G., , Castel, D., , Estelle, J., , Guernec, G., , Jagla, B., , Jouneau, L., , Laloë, D., , Le Gall, C., , Schaëffer, B., , Le Crom, S., , Guedj, M., , Jaffrézic, F., and (2013). A comprehensive evaluation of normalization methods for illumina high-throughput rna sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683.

Friedländer, M. R., MacKowiak, S. D., Li, N., Chen, W., and Rajewsky, N. (2012). MiRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, 40(1):37–52.

Giurato, G., De Filippo, M. R., Rinaldi, A., Hashim, A., Nassa, G., Ravo, M., Rizzo, F., Tarallo, R., and Weisz, A. (2013). IMir: An integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq. *BMC Bioinformatics*, 14:362.

Götz, U., Marker, S., Cheaib, M., Andresen, K., Shrestha, S., Durai, D. A., Nordstrom, K. J., Schulz, M. H., and Simon, M. (2016). Two sets of RNAi components are required for heterochromatin formation in trans triggered by truncated transgenes. *Nucleic Acids Research*, 44:5908–5923.

Han, B. W., Wang, W., Zamore, P. D., and Weng, Z. (2015). 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.

Henras, A. K., Plisson-Chastang, C., O'Donohue, M. F., Chakraborty, A., and Gleizes, P. E. (2015). An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdisciplinary Reviews: RNA*, 6:225–242.

Icay, K., Chen, P., Cervera, A., Rantanen, V., Lehtonen, R., and Hautaniemi, S. (2016). SePIA: RNA and small RNA sequence processing, integration, and analysis. *BioData Mining*, 9:20.

Joh, R. I., Khanduja, J. S., Calvo, I. A., Mistry, M., Palmieri, C. M., Savol, A. J., Ho Sui, S. J., Sadreyev, R. I., Aryee, M. J., and Motamedi, M. (2016). Survival in Quiescence Requires the Euchromatic Deployment of Clr4/SUV39H by Argonaute-Associated Small RNAs. *Molecular Cell*, 64:1088–1101.

Kuksa, P. P., Amlie-Wolf, A., Katanić, ź., Valladares, O., Wang, L. S., and Leung, Y. Y. (2018). SPAR: Small RNA-seq portal for analysis of sequencing experiments. *Nucleic Acids Research*, 46(W1):W36–W42.

Langmead, B., Salzberg, S. L., and Langmead (2013). Bowtie2. *Nature methods*, 9:357–359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Data, G. P., Sam, T., and Subgroup, . G. P. D. P. (2009). The Sequence Alignment / Map format and SAMtools. *Bioinformatics*, 25:2078–2079.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550.

Mohorianu, I., Stocks, M. B., Applegate, C. S., Folkes, L., and Moulton, V. (2017). 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.

Moxon, S., Schwach, F., Dalmay, T., MacLean, D., Studholme, D. J., and Moulton, V. (2008). A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics*, 24(9):2252–2253.

Panero, R., Rinaldi, A., Memoli, D., Nassa, G., Ravo, M., Rizzo, F., Tarallo, R., Milanesi, L., Weisz, A., and Giurato, G. (2017). ISmaRT: A toolkit for a comprehensive analysis of small RNA-Seq data. *Bioinformatics*, 33(16):938–940.

Patwardhan, B., Chaguturu, R., Chavan-Gautam, P., Shah, T., and Joshi, K. (2017). Chapter 8 - Transcriptomics and Epigenomics. In *Innovative Approaches in Drug Discovery*, page 235. Elsevier.

Preer, L. B., Rudman, B., Pollack, S., and Preer, J. R. (1999). 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.

Quek, C., Jung, C. h., Bellingham, S. A., Lonie, A., and Hill, A. F. (2015). iSRAP - A one-touch research tool for rapid profiling of small RNA-seq data. *Journal of Extracellular Vesicles*, 4:29454.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26:841–842.

Rosenkranz, D., Han, C. T., Roovers, E. F., Zischler, H., and Ketting, R. F. (2015). Piwi proteins and piRNAs in mammalian oocytes and early embryos: From sample to sequence. *Genomics Data*, 5:309–313.

Rueda, A., Barturen, G., Lebrón, R., Gómez-Martín, C., Alganza, A., Oliver, J. L., and Hackenberg, M. (2015). SRNAtoolbox: An integrated collection of small RNA research tools. *Nucleic Acids Research*, 43(W1):W467–W473.

Saskói, E., Tibor, K., Sturm, A., Vellai, T., and Weinhardt, N. (2018). Highly efficient RNAi and Cas9-based auto-cloning systems for C. elegans research. *Nucleic Acids Research*, 46(17):e105–e105.

Sultan, M., Schulz, M. H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D., Schmidt, D., O'Keeffe, S., Haas, S., Vingron, M., Lehrach, H., and Yaspo, M.-L. (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, 321(5891):956–960.

Wan, C., Gao, J., Zhang, H., Jiang, X., Zang, Q., Ban, R., Zhang, Y., and Shi, Q. (2017). CPSS 2.0: A computational platform update for the analysis of small RNA sequencing data. *Bioinformatics*, 33(20):3289–3291.

Wang, K., Liang, C., Liu, J., Xiao, H., Huang, S., Xu, J., and Li, F. (2014). Prediction of piRNAs using transposon interaction and a support vector machine. *BMC Bioinformatics*, 15:419.

Wu, J., Liu, Q., Wang, X., Zheng, J., Wang, T., You, M., Sun, Z. S., and Shi, Q. (2013). MirTools 2.0 for non-coding RNA discovery, profiling and functional annotation based on high-throughput sequencing. *RNA Biology*, 10(7):1087–1092.