Automated analysis of small RNA datasets with RAPID

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

April 18, 2018

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.

Abstract

Summary: Understanding the role of small RNA (sRNA) in diverse biological processes is of current interest and often approached through sRNA sequencing. However, analysis of these datasets is difficult due to the complexity of biological RNA processing pathways, which differ between species. Therefore, several different parameters should be analyzed, that were found to guide researchers, such as sRNA strand specificity, length distribution, and distribution of base modifications. We present RAPID, a generic sRNA analysis pipeline, which captures information inherent in the datasets and automatically produces numerous visualizations as user-friendly HTML reports, covering multiple categories required for sRNA analysis. RAPID also facilitates an automated comparison of multiple datasets, with different normalization techniques, and integrates differential expression analysis using DESeq2.
Availability and Implementation: RAPID under MIT license at `https://github.com/SchulzLab/RAPID` or as a bioconda recipe `https://bioconda.github.io/recipes/rapid/README.html`.
Contact: mschulz@mpi-inf.mpg.de

1 Introduction

Widespread availability of small RNA (sRNA) sequencing technologies drives the biological communities in unraveling the pivotal role of sRNA molecules. Although, micro RNA(miRNA)s are the most widely studied sRNA molecules, a growing interest can be seen in other sRNA classes. Understanding sRNA biogenesis and function often involves the computation of various sequence properties like length, strand of origin, and nucleotide modifications of sRNA molecules. While many good sRNA analysis tools exist (Supp. Tab. S1), they focus mainly on predicting novel miRNAs, piwi-interacting RNA (piRNAs), and annotating them. 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.

Hence, we developed a generic sRNA analysis offline tool: Read Alignment, Analysis, and Differential PIpeline (RAPID). 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.

2 Features of RAPID modules

Basic module

The first of four RAPID modules (Fig. 1) is *rapidStats*, which performs sequence (FastQ) alignment, with or without contaminants removal, using Bowtie2 [2]. After alignment, RAPID obtains read statistics such as read length distribution, base modifications, strandedness, and nucleotide content. RAPID can skip the alignment and directly use alignment files (BAM/SAM). To efficiently process, capture and store the aforementioned statistics, RAPID uses SAMtools [3], BEDtools [5], 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. As samples differ in sequencing depth, read counts should be normalized before comparison across samples. Knockdown studies are commonly performed to understand sRNA pathways. In some organisms siRNAs are introduced into the cells to achieve knockdown of genes, but these siRNAs are also sequenced. Thus, they pose an additional challenge for normalization, apart from sequencing depth. RAPID provides two normalization approaches, suitable for different situations (i) DESeq2 based normalization [4] for analysis of many sRNA regions across samples, and (ii) a total count scaling based approach (Eq. (1) in Supp. Methods) that we designed for knockdown-based studies with few regions as input [1].

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 contain 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, base modification, and coverage plots for each gene/region analyzed. In addition, this report also provides combinations of the aforementioned properties. For instance, how is base modification spread 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 any comparative study. 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.

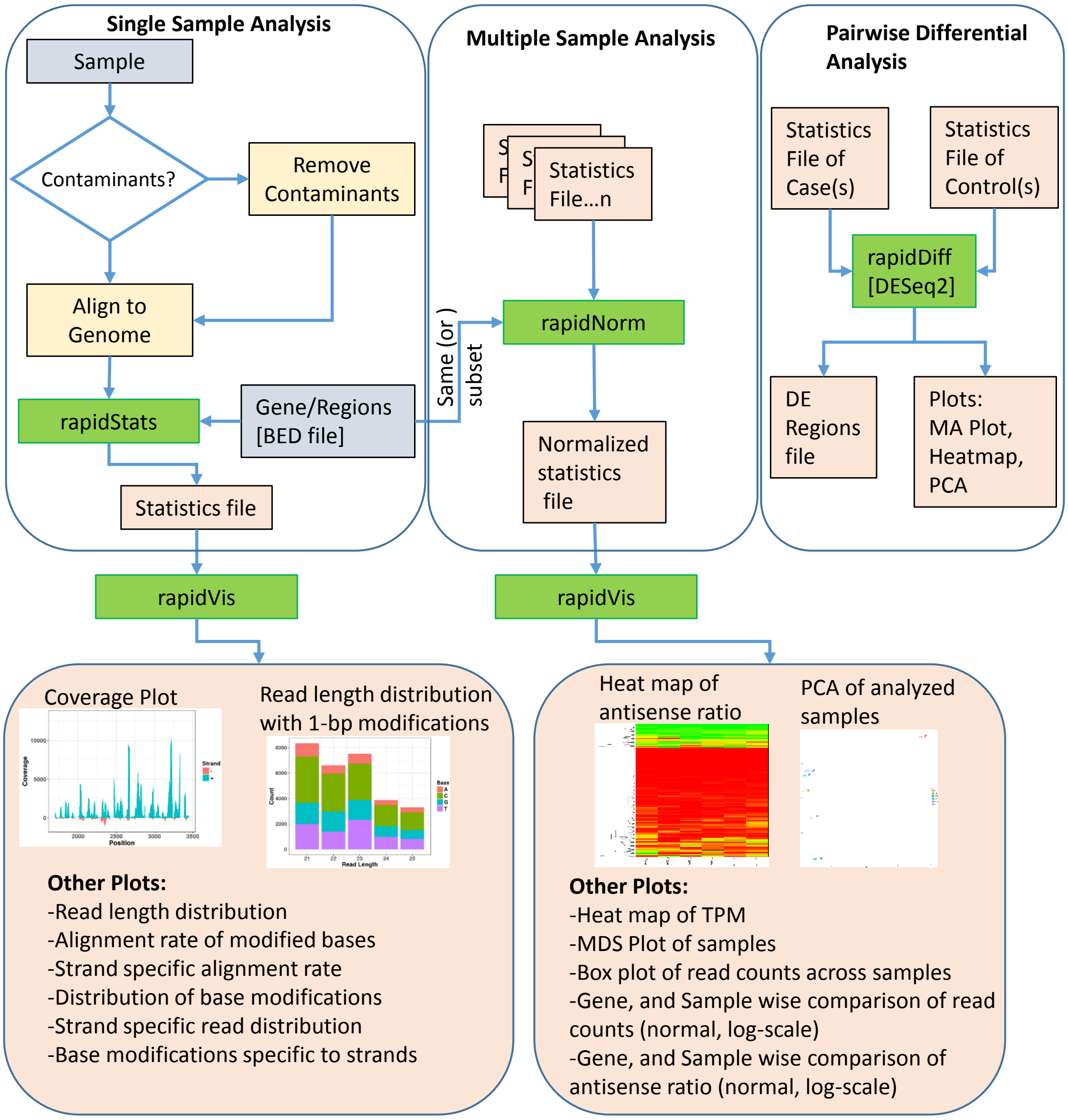
Single Sample Analysis
Sample
Contaminants?
Remove Contaminants
Align to Genome
rapidStats
Gene/Regions [BED file]
Statistics file
Same (or) subset
Multiple Sample Analysis
Statistics File...n
rapidNorm
Normalized statistics file
Pairwise Differential Analysis
Statistics File of Case(s)
Statistics File of Control(s)
rapidDiff [DESeq2]
DE Regions file
Plots: MA Plot, Heatmap, PCA
rapidVis
rapidVis
Coverage Plot
Read length distribution with 1-bp modifications
Other Plots:
-Read length distribution
-Alignment rate of modified bases
-Strand specific alignment rate
-Distribution of base modifications
-Strand specific read distribution
-Base modifications specific to strands
Heat map of antisense ratio
PCA of analyzed samples
Other Plots:
-Heat map of TPM
-MDS Plot of samples
-Box plot of read counts across 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 use case is shown in the boxes. (iv) *rapidDiff* is the differential expression analysis module implementing DESeq2.

Results of the DE analysis include intuitive plots (such as MA Plot, Heatmap, PCA) and the list of DE genes/regions.

3 Case studies

The first example is an analysis of four sRNA-seq datasets (ENA:PRJEB25903) from wild-type serotypes of *P. tetraurelia*. We were interested in sRNAs produced in the rDNA cluster producing 17S, 5.8S, 25S ribosomal RNAs in that organism. As Supp. Figs. S1-S2 show, RAPID analysis reveals an accumulation of 23nt antisense sRNA . Our data here suggests that in *Paramecium* these elimination processes are associated with antisense siRNAs, possibly produced from RNA-dependent RNA Polymerase activity. More details on the analysis and an additional use case on knockdown data from *S. pombe* can be found in the Supplementary Material and Supp. Fig. S3.

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).

Conflict of interest: None declared.

References

[1] 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.

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

[3] 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.

[4] M. I. Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data - the DESeq2 package. *Genome Biology*, 15:550, 2014.

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