shortRNA: A flexible framework for the analysis of short RNA sequencing data

Pierre-Luc Germain[†1,2,3,4], Deepak K. Tanwar[†1,2,3], Isabelle M. Mansuy[1,2]

[1] Institute for Neuroscience, Department of Health Sciences and Technology, ETH Zürich
[2] Laboratory of Neuroepigenetics, Brain Research Inistitute, Medical Faculty of the University of Zürich and Institute for Neuroscience, Department of Health Sciences and Technology of the ETH Zürich
[3] Statistical Bioinformatics Group, Swiss Institute of Bioinformatics
[4] Statistical Bioinformatics Group, University of Zürich

Introduction

Short RNAs, categorized as non-coding RNA molecules, are less than 200 nucleotides in length and play a vital role in the regulation of the genome. They have been implicated in a large number of biological processes and pathological conditions. With the advancement in high-throughput sequencing (HTS) techniques, it is now possible to sequence and study short RNAs[1,2]. However, short RNA studies are hampered by several technical issues, including the bioinformatic analysis of short RNA sequencing data:

1. Analysis methods are not simultaneously optimized towards all known short RNA types (e.g. miRNA, piRNA, rasiRNA, siRNA, snoRNA, tsRNA, tRFs, srRNA and U-RNA). On top of multiplying the work needed for an extensive analysis of the data, this can potentially create misassignment mistakes.

2. Current methods either do not deal adequately with post-transcriptional modifications (for genome-based methods); or if they do (transcript-based methods), they do not deal with unannotated features.

3. Current methods do not adequately account for the hierarchical organization of the features one might want to quantify or test.

4. There is still no consensus on the most appropriate normalization method for short-RNA-Seq data.

Example 1

The issue of post-transcriptional modifications:

tRNA-iMet-CAT-4 is transcribed from chrX, and as other tRNAs receives post-transcriptionally the 3' addition of the nucleotides CCA:

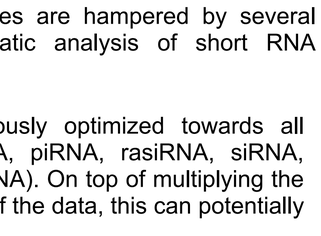
… TCCTCTGCTT → aligns to multiple locations on the genome

↓

… TCCTCTGCTT**CCA** → does not align to the genome

To deal with this, researchers have often simply trimmed 3' CCAs before alignment, however, this can often result in the read becoming ambiguous when instead it initially wasn't, as in the case above. One solution to this issue is to build a custom genome that is complemented with known, post-transcriptionally modified transcripts[3].

Example 2



The issue of ambiguity between related features: tRNA typically have several copies across the genome - for instance, tRNA-Ala-AGC has 23 nearly-identical copies. While a genome alignment will make most reads from such features ambiguous (i.e. multi-mapping), from a functional point of view it is irrelevant from which exact location they came from. This issue becomes even more critical with tRNA fragments, which often have conserved sequences across different tRNAs. One way to address this issue is to aggregate reads into functional equivalence classes, i.e. higher level than specific genes/transcripts.

Objectives

Because of these shortcomings, we developed a new analysis framework that addresses these issues using alternative nested equivalence classes over a customized annotation. We present this approach and package, and show how it can be used to redress biases in the quantification of both specific RNA as well as large RNA classes.

Methods

shortRNA R package is an extension to the **TreeSummarised-Experiment** object.

Database sources

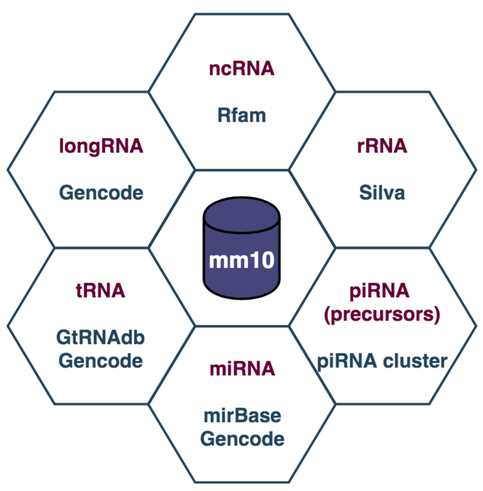


Figure 1: Databases used for mouse.

Tree structure

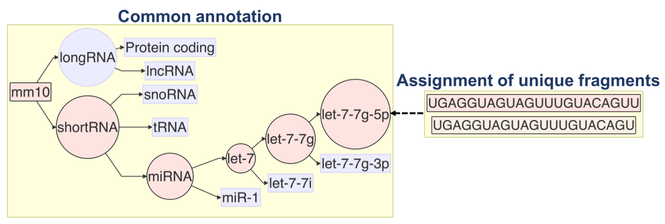


Figure 2: An example tree structure from miRNAs.

Reads assignment problem

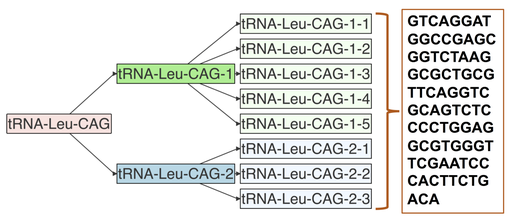


Figure 3: It is hard to assign the reads when tRNAs have same sequence.

Pipeline for short RNA-Seq data analysis

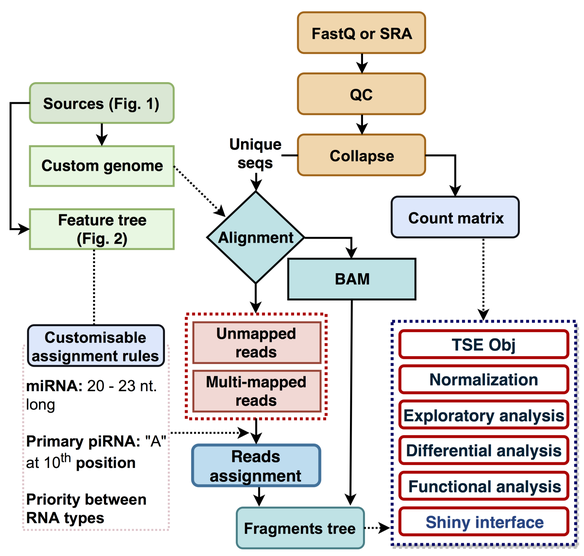


Figure 4: Pipeline for data analysis.

Results

Tree-based reads assignment recovers a large fraction of reads considered ambiguous from a mapping point of view

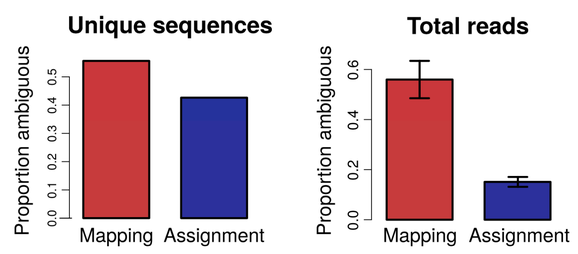


Figure 5: Reads assignment (Sperm short RNA-Seq[4]).

Transcripts abundance plot for uniquely mapping fragments

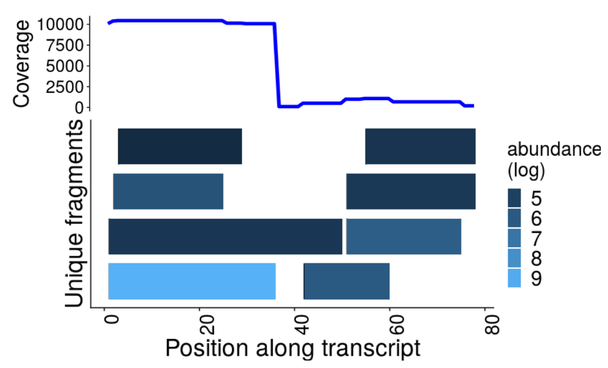


Figure 6: Unique fragments abundance for a transcript (example).

RNA biotype and their proportions

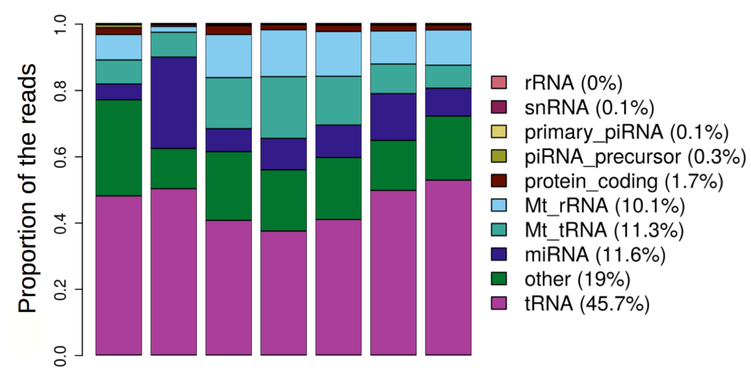


Figure 7: Read proportions from samples (Sperm short RNA-Seq[4]).

Comparison of shortRNA with other methods

We compared 20 published tools for 23 features.

Features	ncPro-Seq	SPORTS 1.0	exceRpt	RAPID	shortRNA
Languages	R, Perl, Python	Perl, R	Java, R	R, Bash, Perl	R
Quality control	✔	✔	✔	✔	✔
Alignment/ Mapping	Bowtie	✔	✔	✔	✔
Databases	miRBase, Rfam, RepeatMasker, User defined	miRBase, rRNA (NCBI), GtRNAdb, piRNA, Ensembl, Rfam	Gencode, mirBase, GtRNAdb, circBase	User defined	Gencode, GtRNAdb, miRbase, piRNA precursors, User defined
Normalization	✘	✘	✔	✔	✔
Differential analysis	✘	✘	✔	✔	✔
Functional analysis	✘	✘	NA	✘	✔
Exploratory data analysis	✔	✔	✔	✔	✔
Adequate handling of post-transcriptional modifications	✘	✔	✘	✘	✔
Unannotated transcripts/ Novel predictions	✔	✔	✔	✘	✔
Heirarchical	✘	✘	✘	✘	✔
isoMirs	✘	✘	✘	✘	✔
User interface	CLI, GUI	CLI	CLI, GUI	CLI	CLI, GUI
Implementation	Tool, Webserver	Tool	Tool, Webserver, Docker	Tool, Conda	R package

Conclusions

- **Standalone** R package for short RNA-Seq data analysis.
- Extension to the **TreeSummarizedExperiment** object.
- QC, alignment, differential analysis, and functional analysis within the package.
- **Hierarchical**: could easily be extended to additional feature trees (for example Vault RNAs).
- shortRNA can be successfully implemented to **investigate the role of short RNAs** in various studies.

Code

Github: mansuylab/shortRNA

References

1. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: The teenage years. Nature Reviews Genetics (2019). doi:10.1038/s41576-019-0150-2

2. Mehta, J. P. Sequencing small RNA: Introduction and data analysis fundamentals. in RNA mapping 93–103 (Springer New York, 2014). doi:10.1007/978-1-4939-1062-5_9

3. Hoffmann, A. et al. Accurate mapping of tRNA reads. Bioinformatics **34**, 1116–1124 (2017).

4. Gapp, K. et al. Implication of sperm RNAs in transgenerational inheritance of the effects of early trauma in mice. Nature Neuroscience **17**, 667–669 (2014).