SOFTWARE TOOL ARTICLE

REVISED Visualization of the small RNA transcriptome using seqclusterViz [version 2; peer review: 2 approved]

seqclusterViz is a stand-alone HTML page that allows the visualization and inspection of small RNA-seq data analyzed with seqcluster tool.

Lorena Pantano [1], Francisco Pantano[2], Eulalia Marti[3], Shannan Ho Sui[1]

[1]Biostatistics, Harvard T.H. Chan school of Public Health, Boston, MA, 02115, USA
[2]Independent Researcher, San Juan, Argentina
[3]Biomedicine, University of Barcelona, Barcelona, Barcelona, Spain

v2 **First published:** 28 Feb 2019, **8**(ISCB Comm J):232 (https://doi.org/10.12688/f1000research.18142.1)
Latest published: 27 Jun 2019, **8**(ISCB Comm J):232 (https://doi.org/10.12688/f1000research.18142.2)

Abstract
The study of small RNAs provides us with a deeper understanding of the complexity of gene regulation within cells. Of the different types of small RNAs, the most important in mammals are miRNA, tRNA fragments and piRNAs. Using small RNA-seq analysis, we can study all small RNA types simultaneously, with the potential to detect novel small RNA types. We describe SeqclusterViz, an interactive HTML-javascript webpage for visualizing small noncoding RNAs (small RNAs) detected by Seqcluster. The SeqclusterViz tool allows users to visualize known and novel small RNA types in model or non-model organisms, and to select small RNA candidates for further validation. SeqclusterViz is divided into three panels: i) query-ready tables showing detected small RNA clusters and their genomic locations, ii) the expression profile over the precursor for all the samples together with RNA secondary structures, and iii) the mostly highly expressed sequences. Here, we show the capabilities of the visualization tool and its validation using human brain samples from patients with Parkinson's disease.

Keywords
small RNA, miRNA, tRNA, snoRNA, sequencing, visualization, report

This article is included in the International Society for Computational Biology Community Journal gateway.

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 2 (revision) 27 Jun 2019		
version 1 28 Feb 2019	✓ report	✓ report

1 **Xavier Bofill-De Ros**, National Cancer Institute, Frederick, USA
National Cancer Institute, Frederick, USA

2 **Stefan Scholten**, Georg-August-University Göttingen, Göttingen, Germany

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Lorena Pantano (lpantano@iscb.org)

Author roles: Pantano L: Conceptualization, Methodology, Software, Supervision, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Pantano F**: Software, Visualization; **Marti E**: Conceptualization, Validation, Visualization; **Ho Sui S**: Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Pantano L, Pantano F, Marti E and Ho Sui S. **Visualization of the small RNA transcriptome using seqclusterViz [version 2; peer review: 2 approved]** F1000Research 2019, **8**(ISCB Comm J):232 (https://doi.org/10.12688/f1000research.18142.2)

First published: 28 Feb 2019, **8**(ISCB Comm J):232 (https://doi.org/10.12688/f1000research.18142.1)

REVISED **Amendments from Version 1**

1. We added how the secondary structure was calculated and the citation to the RNAfold tool
2. We added the precursor size to the locus panel A-right
3. We added help explaining how to modify the abundance profile in panel B-left
4. We changed the sequences abundances to show normalized values in panel C
5. We updated Figure 1 to show the previous changes and increased the resolution

See referee reports

Introduction

Small RNAs are 18-36-nt-long RNA molecules that are involved in gene regulation, chromatin structure, and transposon element repression. The most well known small RNAs are miRNAs, endo-siRNAs and piRNAs[1]. They are typically processed from double-stranded RNA molecules or single-stranded RNA molecules with a hairpin structure[2]. They bind to members of the Argonaute (AGO) protein family to form the RNA-induced silencing complex that regulates other RNA molecules and plays a key role in gene silencing[3,4]. Small RNAs can also regulate chromatin states through histone modification and methylation[5,6]. Next generation sequencing technologies have enabled a deeper understanding of miRNAs, and other small RNA types have been detected. For instance, it is now known that miRNA genes generate several mature variants called isomiRs that have been detected in multiple conditions, tissues and species[7]. Other small RNAs can arise from mature tRNAs (tRNA fragments) or small nucleolar RNAs[8,9]. While the biogenesis of these molecules is not well understood, studies suggest that they bind to AGO proteins and perform similar functions[10,11].

High-throughput sequencing is a powerful technique for detecting and quantifying small RNAs. The analysis of small RNA data involves multiple steps for detection, annotation, quantification, and *de novo* discovery of putative small RNA molecules. In general, tools focus on the annotation of known miRNAs[12], but new methods to detect other functional types of small RNAs are becoming increasingly important to understand the complex roles of small RNAs. Some tools have been developed to address this challenge[13–15] but few of them produce a visual and interactive report[16,17], and many depend on the use of a remote web server[18–21].

We previously developed `seqcluster`, a genome-wide small RNA characterization tool that detects units of transcripts (clusters) using a heuristic iterative algorithm to deal with multi-mapped events[22]. It quantifies all types of small RNAs in non-redundant manner, and extracts patterns of expression in biologically defined groups. This allows us to study any small RNA cluster detected in the samples, including novel regions not previously discovered or small RNAs in species with poorly curated annotations. Here we describe `seqclusterViz`[23], an interactive web-app that reports the output of `seqcluster`, visualizing small RNA biological features to better understand their putative functions. It allows the user to browse lists of detected small RNAs, shows the precursor secondary structures and the small RNA expression on the precursor, allowing for more in-depth characterization of isomiRs, tRNA fragments, and any other small RNAs detected.

`seqcluster` and `seqclusterViz` are integrated into bcbio-nextgen, a community-based Python framework for fully automated high throughput sequencing analysis.

Methods

Implementation

`seqclusterViz`[23] is developed in HTML, CSS and JavaScript programming languages. It is a stand-alone tool without external dependencies. It runs locally on one's computer making it portable and independent. It uses an SQLite JavaScript library to load all the information from a file created by the `seqcluster` tool[22].

Operation

`seqclusterViz`[23] works on Opera >44.0, Firefox >52.0 and Chrome >57.0. It requires a `seqcluster` report as input. An Internet connection is not required. The tool can be downloaded from its home page (https://github.com/lpantano/seqclusterViz/archive/master.zip). After extracting the ZIP file content, the user can open the `index.html` file with the desired web browser. The user first clicks the 'UPLOAD' button and then selects the `seqcluster.db` file. Once the data has been uploaded, the top-left panel displays all of the small RNA transcripts detected. Each small RNA transcript is clickable to obtain more information (1A). After selecting a small RNA transcript, the top-right panel shows the genomic locations for that transcript. The middle-left panel displays the abundance profile along the precursor (1B); the middle-right displays the RNA secondary structure (1B); as calculated by `seqcluster` with `RNAfold` and default parameters[24]; and the bottom table shows the top 50 most abundant sequences. This table can be sorted and searched using text queries (1C).

The tool provides a number of formatting options to emphasize differences between groups and/or samples and to customize figures. Figures can be exported by right-clicking on it. This provides an easy and quick option to generate publication-ready material.

Use cases

We used public data from 14 human brain samples at pre-motor (PT) and motor (CT) stages of Parkinson's disease (GEO accession number GSE97285) and 14 healthy human brain samples (pre-motor controls - PC and motor stages control - CC)[22]. Data was analyzed with `bcbio-nextgen` using `piDNA` to detect the adapter[25], `cutadapt` to remove it[26], `STAR` to align against the hg19 genome assembly[27], and `seqcluster` to detect small RNA transcripts[22]. We used the output `seqcluster.db` from *seqcluster report* command to test `seqclusterViz`[23]. It took four seconds to upload this 28 MB file to the web page. This

Figure 1. `seqclusterViz` features. (**A**) Top panel with table showing the list of small RNAs detected (left) and genomic location (right). (**B**) Middle panel shows abundance profile over the precursor (left), and secondary structure (right). This is an example of batch effect at the 3' end (blue higher than brown) and disease effect at the 5' end (solid lines higher than dashed lines). (**C**) Bottom panel shows a table with the top most expressed sequences on the selected small RNA transcript. The index column is the sequence identifier that links the results to the original seqcluster output files.

dataset is affected by a batch effect for the two Parkinson's groups due to the groups being sequenced at different read lengths. PC and CC samples were derived from the same RNA extraction, and were expected to show similar expression profiles. However, there is a clear difference by batch (brown versus blue) that is visually apparent in the abundance profile of the tRNA-Arg-TCT RNA across the length of the transcript in (1B). Longer reads allow for detection of longer small RNAs since the 3' adapter can be recognized during the analysis (there is a requirement to include adapter sequences in the `seqcluster` tool). The longer reads from the PC/PT samples (blue) permitted detection of longer small RNAs at the end of the precursor, generating the batch difference in the abundance profile. Moreover, there is a difference in expression at the 5' end of the precursor, where Parkinson's samples (solid lines) are higher than their respective controls (dashed lines). The secondary structure of this small RNA shows a pre-miRNA-like hairpin structure (with a stem-bulge-stem and a terminal-loop) that is normally required to be processed into 18-33-nt mature molecules, where the stem-bulge-stem section encodes the mature sequence[28,29]. Although the structure is larger than typical pre-miRNAs, it is still possible to process with the miRNA machinery. Thus the secondary structure of the molecule can serve as an additional feature to evaluate when seeking candidates for further

experimental validation. Quantitative polymerase chain reaction (qPCR) or small RNA transfection technologies are often used to validate small RNA stability and function. To do so, a single small RNA needs to be used as the target sequence for these assays. The table at the bottom of the page (1C) allows users to select the most abundant sequence in the current small RNA that can be used for such experiments.

Summary

`seqclusterViz`[23] helps users to explore the expression profiles of detected small RNAs across the length of the precursor, the secondary structure of the small RNA, and the annotation. We show the importance of visualizing small RNAseq data to prioritize candidate small RNAs for further experimental validation or functional analysis. The user can modify the figure format and export it for publication or presentation purposes. It is also possible to select the most highly expressed sequence of a transcript cluster that can be used for qPCR or for cell transfection assays.

Data availability

Data to reproduce this analysis is available from the Parkinson project page.

Data from 14 healthy human brain samples were originally reported by Pantano *et al.*[22]. Data from 14 human brain samples at pre-motor (PT) and motor (CT) stages of Parkinson's disease are available at GEO, accession number GSE97285.

The web-tool can be tested at GitHub pages. Click on **Load Example** to start using the tool with the example data set.

Software availability

`seqclusterViz` **is downloaded from:** https://github.com/lpantano/seqclusterViz/archive/v0.1.2.zip.

Source code available from: https://github.com/lpantano/seqclusterViz.

Link to source code as at time of publication: url https://doi.org/10.5281/zenodo.3250205[23].

License: MIT License.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgments

The authors would like to thank researchers who helped to improve this tool: Aron Gyuris, Mira Pavkovic, Maria Mavrikaki. Thank you also to Amanda King for edits.

References

1. Martens-Uzunova ES, Olvedy M, Jenster G: **Beyond microRNA--novel RNAs derived from small non-coding RNA and their implication in cancer.** *Cancer Lett.* 2013; **340**(2): 201–211.
PubMed Abstract | Publisher Full Text
2. Kim VN, Han J, Siomi MC: **Biogenesis of small RNAs in animals.** *Nat Rev Mol Cell Biol.* 2009; **10**(2): 126–139.
PubMed Abstract | Publisher Full Text
3. Kim DH, Saetrom P, Snøve O Jr, *et al.*: **MicroRNA-directed transcriptional gene silencing in mammalian cells.** *Proc Natl Acad Sci U S A.* 2008; **105**(42): 16230–16235.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Okamura K, Lai EC: **Endogenous small interfering RNAs in animals.** *Nat Rev Mol Cell Biol.* 2008; **9**(9): 673–678.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Moazed D: **Small RNAs in transcriptional gene silencing and genome defence.** *Nature.* 2009; **457**(7228): 413–420.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Gonzalez S, Pisano DG, Serrano M: **Mechanistic principles of chromatin remodeling guided by siRNAs and miRNAs.** *Cell Cycle.* 2008; **7**(16): 2601–2608.
PubMed Abstract | Publisher Full Text
7. Zhang Y, Zang Q, Xu B, *et al.*: **IsomiR Bank: a research resource for tracking IsomiRs.** *Bioinformatics.* 2016; **32**(13): 2069–2071.
PubMed Abstract | Publisher Full Text
8. Kawaji H, Nakamura M, Takahashi Y, *et al.*: **Hidden layers of human small RNAs.** *BMC Genomics.* 2008; **9**(1): 157.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Telonis AG, Loher P, Honda S, *et al.*: **Dissecting tRNA-derived fragment complexities using personalized transcriptomes reveals novel fragment classes and unexpected dependencies.** *Oncotarget.* 2015; **6**(28): 24797–822.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Cole C, Sobala A, Lu C, *et al.*: **Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs.** *RNA.* 2009; **15**(12): 2147–2160.
PubMed Abstract | Publisher Full Text | Free Full Text
11. Brameier M, Herwig A, Reinhardt R, *et al.*: **Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs.** *Nucleic Acids Res.* 2011; **39**(2): 675–686.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Lukasik A, Wójcikowski M, Zielenkiewicz P: **Tools4miRs - one place to gather all the tools for miRNA analysis.** *Bioinformatics.* 2016; **32**(17): 2722–4.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Baras AS, Mitchell CJ, Myers JR, *et al.*: **miRge - A Multiplexed Method of Processing Small RNA-Seq Data to Determine MicroRNA Entropy.** *PLoS One.* 2015; **10**(11): e0143066.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Beckers M, Mohorianu I, Stocks M, *et al.*: **Comprehensive processing of high-throughput small RNA sequencing data including quality checking, normalization, and differential expression analysis using the UEA sRNA Workbench.** *RNA.* 2017; **23**(6): 823–835.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Giurato G, De Filippo MR, Rinaldi A, *et al.*: **iMir: an integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq.** *BMC Bioinformatics.* 2013; **14**: 362.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Stocks MB, Moxon S, Mapleson D, *et al.*: **The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets.** *Bioinformatics.* 2012; **28**(15): 2059–2061.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Quek C, Jung CH, Bellingham SA, *et al.*: **iSRAP - a one-touch research tool for rapid profiling of small RNA-seq data.** *J Extracell Vesicles.* 2015; **4**: 29454.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Rueda A, Barturen G, Lebrón R, *et al.*: **sRNAtoolbox: an integrated collection of small RNA research tools.** *Nucleic Acids Res.* 2015; **43**(W1): W467–73.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Zhang Y, Xu B, Yang Y, *et al.*: **CPSS: a computational platform for the analysis of small RNA deep sequencing data.** *Bioinformatics.* 2012; **28**(14): 1925–1927.
PubMed Abstract | Publisher Full Text

20. Yang JH, Qu LH: **DeepBase: annotation and discovery of microRNAs and other noncoding RNAs from deep-sequencing data.** *Methods Mol Biol.* 2012; **822**: 233–248.
PubMed Abstract | Publisher Full Text

21. Huang PJ, Liu YC, Lee CC, *et al.*: **DSAP: deep-sequencing small RNA analysis pipeline.** *Nucleic Acids Res.* 2010; **38**(Web Server issue): W385–91.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Pantano L, Friedländer MR, Escaramís G, *et al.*: **Specific small-RNA signatures in the amygdala at premotor and motor stages of Parkinson's disease revealed by deep sequencing analysis.** *Bioinformatics.* 2016; **32**(5): 673–681.
PubMed Abstract | Publisher Full Text

23. Pantano L, franpantano: **lpantano/seqclusterviz: v0.1.2.** 2019.
http://www.doi.org/10.5281/zenodo.3250205

24. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, *et al.*: **Viennarna Package 2.0.** *Algorithms Mol Biol.* 2011; **6**(1): 26.
PubMed Abstract | Publisher Full Text | Free Full Text

25. Tsuji J, Weng Z: **DNApi: A *De Novo* Adapter Prediction Algorithm for Small RNA Sequencing Data.** *PLoS One.* 2016; **11**(10): e0164228.
PubMed Abstract | Publisher Full Text | Free Full Text

26. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads.** *EMBnet J.* 2011; **17**(1): 10.
Publisher Full Text

27. Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text | Free Full Text

28. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell.* 2004; **116**(2): 281–97.
PubMed Abstract | Publisher Full Text

29. Feng Y, Zhang X, Graves P, *et al.*: **A comprehensive analysis of precursor microRNA cleavage by human Dicer.** *RNA.* 2012; **18**(11): 2083–92.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 01 April 2019

https://doi.org/10.5256/f1000research.19842.r45091

✔ **Stefan Scholten**
Crop Plant Genetics, Georg-August-University Göttingen, Göttingen, Germany

The article reports on a visualization tool for seqcluster outputs, a software tool to characterise small transcriptome data. It is a smart tool for sRNA visualization. The interactive view makes it attractive, especially the visualization of secondary structures. The performance of the tool fits description and the filter option is very helpful for jumping to the desired information.

The tool is restricted to seqcluster.db file as input and cannot be used as a general-purpose sRNA visulization tool using map files.

Sufficient information is provided to allow interpretation of the expected output data sets and results generated with the tool. The example expression profile provides clear distinction between samples.

The editing option makes it even better, but when one switches to the line view different lengths can be seen. It is not clear whether this is related with the length of the sRNA that map at that position. A summary of the lengths that mapped would be useful additional information. Alongside the description section in Figure A (left side) information on the length of sRNA should be included.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Transcriptomics, developmental biology, reproduction

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 10 Apr 2019

Lorena Pantano Rubino, Harvard T.H. Chan school of Public Health, Boston, USA

Dear Stefan Scholten,

Thanks for finding time to review this article. I will work on your recommendation to add that information to the tool, it is a very good idea we missed.

I'll post back the update version here.

Cheers

Competing Interests: No competing interests were disclosed.

Reviewer Report 18 March 2019

https://doi.org/10.5256/f1000research.19842.r45803

Xavier Bofill-De Ros
[1] RNA Mediated Gene Regulation Section, RNA Biology Laboratory, Center for Cancer Research, National Cancer Institute, Frederick, MD, USA
[2] National Cancer Institute, Frederick, USA

This article by *Pantano et al.* describes a novel software, SeqclusterViz, to visualize and help the interpretation of small RNA-seq data mapped using Seqcluster. Moreover, the authors illustrate its usage with a case study of small RNAs dysregulated in the brain of patients with Parkinson's disease. This software is of importance to help researchers do in depth analysis and compare the exact mapping of reads across multiple samples. The subject of this manuscript is interesting and seems adequate for indexing in F1000Research. The work looks solid although minor edits could be performed to improve the

quality and usability of the SeqclusterViz.

Minor comments:

- The SeqclusterViz displays the sequence of the reads mapped to a precursor RNA such as a pri-miRNA. In particular, mature miRNA is well known for having isoforms (isomiRs) with the addition or internal edit to non-templated nucleotides. The field of isomiR study is of growing interest, SeqclusterViz will benefit from a display of non-templated nucleotides.
- A secondary structure prediction is implemented on SeqclusterViz, please describe in more the prediction method used. Include input parameters (such as if GU pairs at the end of helices) and outputs such as MFE, bracket-dot notation...
- Provide units for the Y and X axis in “Abundance profile along precursor”.
- Provide normalized read counts on “Table with Sequences”.
- Repair the path to make the button “Load example” active.
- The buttons of “Add filter” and “Change line” can’t be linked to the samples easily.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: MicroRNA biogenesis and isoforms

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 10 Apr 2019

Lorena Pantano Rubino, Harvard T.H. Chan school of Public Health, Boston, USA

Dear Xavier Bofill-De Ros,

Thanks a lot to find time to review this article. I will take actions on all the points I can address in the next month and report an update as soon as possible.

Cheers

Competing Interests: No competing interests were disclosed.