Check for updates

SOFTWARE TOOL ARTICLE

***RNA*tor: an Android-based application for biologists to plan RNA sequencing experiments [version 1; peer review: 1 approved with reservations, 1 not approved]**

Shruti Kane¹, Himanshu Garg¹, Neeraja M. Krishnan¹, Aditya Singh¹,
Binay Panda ^{1,2}

¹Ganit Labs, Bio-IT Centre, Institute of Bioinformatics and Applied Biotechnology, Bangalore, India

²Strand Life Sciences, Bangalore, India

V1 First published: 26 Jun 2017, 6:997
<https://doi.org/10.12688/f1000research.11982.1>

Latest published: 16 Nov 2017, 6:997
<https://doi.org/10.12688/f1000research.11982.2>

Abstract

RNA sequencing (RNA-seq) is a powerful technology that identifies novel transcripts (coding, non-coding and splice variants), understands transcript structures, and estimates gene/allele expression. Biologists face specific challenges while designing RNA-seq experiments. The nature of these challenges lies in determining the total number of sequenced reads and replicates required for detecting marginally differentially expressed transcripts, and in determining the adequate number of lanes to use in a sequencing flow cell. Despite previous attempts to address these challenges, easily accessible and biologist-friendly mobile applications do not exist. Thus, we developed *RNA*tor, a mobile application for Android platforms, to aid biologists in correctly designing their RNA-seq experiments. The recommendations from *RNA*tor are based on simulations and real data.

Keywords

RNA-seq, Android-based, simulations, mobile application, recommendations, experimental design

Open Peer Review

Approval Status

	1	2
version 2
(revision)
16 Nov 2017
version 1
26 Jun 2017

1. **Daisuke Komura**, Tokyo Medical and Dental University, Tokyo, Japan

2. **Niranjan Nagarajan**, Genome Institute of Singapore, Singapore, Singapore

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

Corresponding author: Binay Panda (binay@ganitlabs.in)

Author roles: Kane S: Formal Analysis, Methodology; Garg H: Software; Krishnan NM: Formal Analysis, Writing – Review & Editing; Singh A: Formal Analysis, Methodology; Panda B: Conceptualization, Project Administration, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Research presented in this article is funded by the Department of Electronics and Information Technology, Government of India (Ref No: 18(4)/2010-E-Infra., 31-03-2010) and Department of IT, BT and ST, Government of Karnataka, India (Ref No: 3451-00-090-2-22).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2017 Kane S *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Kane S, Garg H, Krishnan NM *et al.* ***RNAator: an Android-based application for biologists to plan RNA sequencing experiments [version 1; peer review: 1 approved with reservations, 1 not approved]*** F1000Research 2017, 6:997 <https://doi.org/10.12688/f1000research.11982.1>

First published: 26 Jun 2017, 6:997 <https://doi.org/10.12688/f1000research.11982.1>

Introduction

RNA-seq offers several advantages over low-throughput technologies such as quantitative PCR and annotation-dependent methods such as microarrays. Designing RNA-seq experiments accurately, however, poses challenge to biologists. This is particularly true when prior knowledge on genome or transcriptome of the organism of choice is not available. It is important to determine the number of replicates and the number of sequencing reads, and choose the right analytical tool, to estimate subtle differences between expression levels of transcripts. In the current manuscript, we describe *RNAator*, an Android app with a user-friendly graphical user interface (GUI) that helps biologists design RNA-seq experiments. *RNAator* can be linked to any existing differential expression analysis tool, and can help design experiments to estimate expression differences of as low as 0.8–1.2X fold. *RNAator's* recommendations are based on an exhaustive combination of discovery with simulated reads for transcriptomes of varying sizes (3 to 100 MB), followed by validation with real datasets on wild-type and mutant conditions of *Saccharomyces cerevisiae*.

Methods

Implementation

We simulated varying numbers of Illumina-like reads with replicates, with fold changes ranging from 1.2–5X between the control and treatment samples on a 3 MB human chr14 (hg19) transcriptome, using Polyester (Frazee *et al.*, 2015). We detected differentially expressed genes (DEGs) on all the simulations using Tophat v2.1.1-Cufflinks v2.2.1 (Trapnell *et al.*, 2012) based genome-guided workflow followed by differential expression analyses using five tools: Deseq v1.28.0 (Anders & Huber, 2010); Deseq2 v1.16.1 (Love *et al.*, 2014); EdgeR v3.18.1 (Robinson *et al.*, 2010); Cuffdiff-Cufflinks v2.2.1 (Trapnell *et al.*, 2012); and Kallisto v0.43.1 (Bray *et al.*, 2016) and a *de novo* assembly-based tool, Trinity v2.3.2 (Grabherr *et al.*, 2011) followed by differential expression analyses using Kallisto v0.43.1 (Bray *et al.*, 2016). We studied results from these simulations on the number of DEGs detected reliably and the extent of recovery of those DEGs. Based on these simulations, we arrived at recommendations on the number of reads, number of replicates, and the tool(s) needed to identify DEGs reliably. We validated these recommendations using simulated reads from larger transcriptomes (10MB, 30MB and 100MB), created by combining transcriptomes from more than one hg19 chromosome, and using a real *Saccharomyces cerevisiae* dataset (ENA accession: ERP004763) comprising of 48 biological replicates, for two conditions; wild-type (WT) and a *snf2* knock-out (KO) mutant.

Operation

The size of the transcriptome (or genome if the transcriptome size is not known), taken from a user-defined or from a backend database, the number of replicates to use and the fold change of DEGs are user-defined parameters in *RNAator* (Figure 1). An

RNAator flowchart highlighting simulation conditions and analytical tools used is provided in [Supplementary Figure S1](#).

Results

RNAator was evaluated using questions that a biologist would typically ask before starting an experiment, followed by the recommendations provided by *RNAator*.

Read requirements for optimal DEG detection

1, 1.5, 6, 10, 14 and 20 million reads are needed for detection of differential expression of DEGs at 5-fold, 4-fold, 3-fold, 2-fold, 1.5-fold and 1.2-fold change, respectively, for a 3Mb transcriptome with 3 replicate samples.

We simulated 0.2–20 million reads for human chromosome 14 (~3Mb) and observed that the numbers of detected DEGs simulated at a given fold change peaked for a certain coverage before plateauing (Figure 2). This observation remained valid for the real data (Figure 3) and the large simulated transcriptomes (10Mb, 30Mb and 100Mb) (Supplementary Figure S2). Increasing the number of sequencing reads increased the sensitivity of detection. The final recommendations from *RNAator* correspond to the number of DEGs at its peak, and are therefore a good compromise between sensitivity and keeping the cost of sequencing low. Changing the number of replicates does change the recommendation. For example, with more than three replicates, *RNAator* suggests producing fewer reads to obtain the same information (Table 1).

Detection sensitivity of DE tools

Kallisto detected optimal number of DEGs with the highest sensitivity. Focusing purely on the number of DEGs detected between WT and KO, Kallisto performed best over the other tools tested (Figure 2 and Supplementary Figure 3).

Detection specificity and transcript recovery by DE tools

Cuffdiff can be used for high specificity and DeSeq2 and EdgeR, for high transcript recovery. Although Kallisto-Sleuth was fast and produced results with high sensitivity; we observed that this was at the expense of specificity of detection (Supplementary Figure S3). Cuffdiff produced results with high specificity albeit with a loss of sensitivity (Supplementary Figure S3). The transcript recovery was best for EdgeR among the 3 tools tested (CuffDiff, DeSeq and EdgeR, Supplementary Figure S4).

Performance of assembly-based pipeline over that of genome-guided tools

The assembly-based pipeline yields more DEGs with higher specificity. Using Trinity (Grabherr *et al.*, 2011) as an assembly pipeline along with Kallisto enhanced the number of DEGs detected when compared with the genome-guided Kallisto-Sleuth pipeline (Figure 2). The specificity of Trinity-Kallisto was also better when compared to the Kallisto-Sleuth pipeline (Supplementary Figure S3).

Figure 1. Screenshots of the *RNAtor* mobile application.

Figure 2. Number of differentially expressed genes (DEGs) detected for simulated datasets (hg19 chr14) by Deseq, Deseq2, EdgeR, Cuffdiff, Kallisto-Sleuth and Trinity-Kallisto tools.

Figure 3. Number of differentially expressed genes (DEGs) detected using a real dataset (*Saccharomyces cerevisiae*) with the Kallisto-Sleuth pipeline.

Table 1. RNAator output on the number of sequencing reads (in millions) to be produced for 2–5 sample replicates to detect differentially expressed genes at a given fold change.

	2 replicates	3 replicates	4 replicates	5 replicates
5fold	6	2	1.5	1.5
4fold	10	6	2	1.5
3fold	10	6	6	6
2fold	14	10	10	6
1.5fold	30	20	20	14

Discussion

Although some of the challenges with RNA-seq experiments have been addressed previously (Busby *et al.*, 2013; Luo *et al.*, 2014), currently there is no easy-to-use, biologist-friendly mobile phone-based app. Scotty, a previously reported, useful, interactive web-based tool aids RNA-seq experimental design. However, it has a dependence on pilot or prototype data, closely matching the actual experimental conditions (Busby *et al.*, 2013). Additionally, it can detect genes or transcripts of only up to 2X fold change in the test condition relative to the control. RNAator addresses some of these gaps as a user-friendly mobile app, however it has certain limitations. For example, it does not take into account the dynamic nature of any transcriptome (where the exact size of transcriptome is not known and cannot simply be derived from the genome size), the throughput of different sequencing instruments and the presence of spliced variants. These limitations will be addressed in future releases.

Data availability

The Android version of RNAator is available on Google Play Store.

Latest source code: <https://github.com/binaypanda/RNAator>.

Archived source code as at the time of publication: <https://doi.org/10.5281/zenodo.814905> (Panda, 2017).

License: RNAator v1.0 is distributed under GNU GPLv3 licence.

Competing interests

No competing interests were disclosed.

Grant information

Research presented in this article is funded by the Department of Electronics and Information Technology, Government of India (Ref No: 18(4)/2010-E-Infra., 31-03-2010) and Department of IT, BT and ST, Government of Karnataka, India (Ref No: 3451-00-090-2-22).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary Figure S1: *RNAator* flowchart highlighting simulation conditions (reads, replicates, and fold change of differential expression) and analytical tools used.

[Click here to access the data.](#)

Supplementary Figure S2: Number of differentially expressed genes (DEGs) detected for various simulated dataset on 10Mb, 30Mb and 100Mb transcriptomes using the Kallisto-Sleuth pipeline.

[Click here to access the data.](#)

Supplementary Figure S3: True/false positive curves for differentially expressed genes (DEGs) recovered under various simulation conditions, created by combining reads (0.1M–20M), replicates (2–5) and fold change of differential expression (1.2–5X) by Cuffdiff, Deseq2, EdgeR, Kallisto and Trinity-Kallisto tools.

[Click here to access the data.](#)

Supplementary Figure S4: Percentage recovery of transcripts under various simulation conditions, created by combining reads (0.1M–20M), replicates (0–5) and fold change of differential expression (1.2–5X) with CuffDiff, DeSeq and EdgeR. The size of the bubble represents the extent of transcript recovery.

[Click here to access the data.](#)

References

Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol.* 2010; **11**(10): R106.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic RNA-seq quantification.** *Nat Biotechnol.* 2016; **34**(5): 525–527.

[PubMed Abstract](#) | [Publisher Full Text](#)

Busby MA, Stewart C, Miller CA, *et al.*: **Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression.** *Bioinformatics.* 2013; **29**(5): 656–657.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Frazee AC, Jaffe AE, Langmead B, *et al.*: **Polyester: simulating RNA-seq datasets with differential transcript expression.** *Bioinformatics.* 2015; **31**(17): 2778–2784.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Grabherr MG, Haas BJ, Yassour M, *et al.*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol.* 2011; **29**(7): 644–652.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

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

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Luo H, Li J, Chia BK, *et al.*: **The importance of study design for detecting differentially abundant features in high-throughput experiments.** *Genome Biol.* 2014; **15**(12): 527.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Panda B: **binaypanda/RNator: RNator.** 2017.

[Data Source](#)

Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**(1): 139–140.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Trapnell C, Roberts A, Goff L, *et al.*: **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.** *Nat Protoc.* 2012; **7**(3): 562–578.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status: ? ✖

Version 1

Reviewer Report 30 August 2017

<https://doi.org/10.5256/f1000research.12955.r23782>

© 2017 Nagarajan N. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Niranjan Nagarajan

Genome Institute of Singapore, Singapore, Singapore

RNAator looks like a simple and easy to use application. However this could also be its drawback in that it may be too simplistic. The authors should show some evidence that the parts they omit do not significantly impact RNAator's guidance to users.

Abstract:

- 1) Perhaps the first sentence of the abstract should be reworded as it is a bit odd to say that "RNA Sequencing ... understands transcript structures" etc.
- 2) Its unnecessary to mention the "number of lanes" when the "number of reads" has been highlighted.
- 3) It is s not clear why a mobile application is necessary. Is a web-based tool like EDDA not sufficient?

Introduction:

- 4) The last sentence is not very clear in terms of what is being done here.
- 5) As it is typical to present and discuss prior work in the introduction, I believe it is important to add this component. Some of this could perhaps come from the current discussion section.

Implementation:

- 6) What were the parameters used for running Polyester? In particular, different experimental conditions will likely have different expression profiles, biological variability dictating the dispersion parameters and perhaps other factors like splicing complexity, gene families etc. that all affect RNA-seq analysis. How does RNAator account for these? How were the various differential analysis softwares run? They usually have cutoffs that can be chosen to make their results more or

less stringent.

Results:

7) The detection of differentially abundant RNAs using RNA-seq is impacted by the relative abundance of the transcript. How is this taken into account in table 1?

8) In figure 2, are these all true positives? What about false positives? Usually there is a tradeoff and a user needs to be aware of this.

9) Figure 3 does not have a legend.

10) It is not clear to me what transcript recovery is referring to and why that is relevant here.

11) Supplementary figure 4 needs better resolution and font sizes.

12) The claims made in the section "Detection specificity and transcript recovery by DE tools" are not obvious from the figures shown.

Discussion:

13) It seems to me that there are many more limitations in RNAtor than are discussed here. Also, it is appropriate to discuss the strengths and weaknesses of EDDA as well and perhaps some of this should be in the introduction, as noted earlier.

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

No

Is the description of the software tool technically sound?

No

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

Partly

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?

No

Competing Interests: My lab developed the software EDDA (<http://edda.gis.a->

star.edu.sg/; <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0527-7>) which has partly overlapping functionality.

Reviewer Expertise: Genomics, Computational Biology

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 10 Nov 2017

Binay Panda, Institute of Bioinformatics and Applied Biotechnology, Bangalore, India

We thank Dr. Niranjana Nagarajan for reviewing the manuscript and providing his comments. Following his suggestions, we have revised the manuscript and have responded to all his queries below. We believe this has improved the manuscript.

Reviewer's comments:

RNAator looks like a simple and easy to use application. However this could also be its drawback in that it may be too simplistic. The authors should show some evidence that the parts they omit do not significantly impact RNAator's guidance to users.

Authors' response:

We agree with the reviewer that the RNAator v1.0 is simplistic and some of the assumptions can affect the recommendations. However, the tool is intended to be used by biologists who use RNA-seq as a replacement to existing technology like microarray and qPCR, to ask simpler biological questions. Many small biology labs, we believe, are still in this group. The tool is not intended to serve advanced users. We are well aware of its limitations and intend to expand the scope of the tool in its future versions, by introducing biases that mimic various experimental conditions into the simulation phase. Nevertheless, we believe that the validation of the recommendations resulting from training on simulated RNA-seq data that has not yet incorporated various biological biases, with real data from *Saccharomyces cerevisiae* provides strong evidence that our assumptions do not significantly impact RNAator's guidance to users.

Abstract:

Reviewer's comments:

1) Perhaps the first sentence of the abstract should be reworded as it is a bit odd to say that "RNA Sequencing ... understands transcript structures" etc.

Authors' response:

We thank the reviewer to point this. We have re-worded this sentence in the revised

manuscript.

Reviewer's comments:

2) Its unnecessary to mention the "number of lanes" when the "number of reads" has been highlighted.

Authors' response:

Following the reviewer's suggestion, we have removed the use of the term in the revised manuscript.

Reviewer's comments:

3) It is not clear why a mobile application is necessary. Is a web-based tool like EDDA not sufficient?

Authors' response:

We agree with the reviewer that there are web-based tools, like EDDA and Scotty, which we have now elaborated in the revised manuscript, that can do similar job. However, we strongly believe that a mobile application offers a lot more flexibility, ease of navigation, and user-friendliness, compared to a web-based tool. Besides, many features of the App are available for offline use and the user has the flexibility to use it on the work bench or anywhere convenient. Additionally, Apps use a different framework to run code and hence, run faster than mobile websites that typically use javascript to run their code, and store the data locally on the mobile device, unlike websites that store data on web servers.

Introduction:

Reviewer's comments:

4) The last sentence is not very clear in terms of what is being done here.

Authors' response:

Following the reviewer's suggestion, we have modified the sentence in the revised manuscript.

Reviewer's comments:

5) As it is typical to present and discuss prior work in the introduction, I believe it is important to add this component. Some of this could perhaps come from the current discussion section.

Authors' response:

We have now modified the introduction to include references on tools aiding RNA-seq experimental design.

Implementation:

Reviewer's comments:

6) What were the parameters used for running Polyester? In particular, different experimental conditions will likely have different expression profiles, biological variability dictating the dispersion parameters and perhaps other factors like splicing complexity, gene families etc. that all affect RNA-seq analysis. How does RNAtor account for these? How were the various differential analysis softwares run? They usually have cutoffs that can be chosen to make their results more or less stringent.

Authors' response:

The per_reads_transcript, num_reps, fold_changes, in addition to the input 'fasta' and output 'outdir' parameters were exercised in polyester simulations. All differential expression analysis softwares were run with default cut-offs.

We have now mentioned these in the revised manuscript

Results:

Reviewer's comments:

7) The detection of differentially abundant RNAs using RNA-seq is impacted by the relative abundance of the transcript. How is this taken into account in table 1?

Authors' response:

In the current version of the tool implementation, the simulated data does not take into account the relative abundance of transcript, for e.g. in relation to a control gene or any other gene of interest. Hence, our recommendations (Table 1) are not affected by these. However, differences in relative abundances of true control genes between treatment and control would lead to over- or under-estimation of DEGs, and therefore, the base-line assumption of $\log_2 FC=0$ should be accordingly adjusted.

We have acknowledged this limitation in the revised manuscript.

Reviewer's comments:

8) In figure 2, are these all true positives? What about false positives? Usually there is a tradeoff and a user needs to be aware of this.

Authors' response:

Figure 2 are all detected DEGs, *i.e.*, both true and false positives. Supplementary Figure S3 gives the true positive and false positive trends separately. Yes, there is a tradeoff between sensitivity (true positives) and specificity (absence of false positives). Kallisto performs with a high sensitivity while compromising on specificity, while CuffDiff performs with a high specificity while compromising on sensitivity. We have elaborated this in the revised manuscript.

Reviewer's comments:

9) Figure 3 does not have a legend.

Authors' response:

We thank the reviewer to point this out and have now added a legend to Figure 3.

Reviewer's comments:

10) It is not clear to me what transcript recovery is referring to and why that is relevant here.

Authors' response:

Transcript recovery refers to the length of the transcript as assembled by Tophat, detected as differentially expressed by EdgeR or CuffDiff or DESeq2, in relation to the actual length, as per simulations. It is possible to estimate this parameter only for these three tools, since they offer a handle to the actual transcript Ids. This has been clarified in the revised manuscript.

Reviewer's comments:

11) Supplementary figure 4 needs better resolution and font sizes.

Authors' response:

Supplementary Figure 4 with improved resolution has now been added.

Reviewer's comments:

12) The claims made in the section "Detection specificity and transcript recovery by DE tools" are not obvious from the figures shown.

Authors' response:

The results in this section have been elaborated further with reference to Supplementary Figure S4 in the revised manuscript. With reference to the claim made about detection specificity citing Supplementary Figure S3, we found that CuffDiff performs with high specificity at the loss of sensitivity, with the opposite being true for Kallisto-Sleuth.

Competing Interests: No competing interests were disclosed.

Reviewer Report 06 July 2017

<https://doi.org/10.5256/f1000research.12955.r24054>

© 2017 Komura D. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Daisuke Komura

Department of Genomic Pathology, Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan

The authors developed a new mobile application named RNAtor to assist in the designing of RNA-seq experiments. It provides users with the number of reads that is required for optimal detection of differentially expressed genes at a given fold-change threshold based on simulations and real data.

This is a useful tool for NGS users. However there are some errors and suggestions that need to be fixed.

Major:

1. In general, some reads in RNA-seq analysis are removed due to their low sequencing quality or cannot be mapped probably due to library contamination of rRNA or other species' RNA. Does this software consider the effect? If not, it could underestimate the number of required reads.
2. How does RNAtor calculate the number of reads required for DEG detection? The authors claim that it is calculated based on the number of DEGs at its peak in both real and simulated data but the exact algorithm is not shown. Specifically, how were the peak values calculated and how were the peak values of real and simulated data merged?

Minor:

1. Which does the term 'replicates' mean in this manuscript, technical replicate or biological replicate?
2. Implementation: "... workflow followed by differential expression analysis using five tools:..." I think four instead of five is correct because Kallisto does not use outputs of Tophat-Cufflinks pipeline (Supplementary Figure 1).
3. Figure 2: How many DEGs were included in total in the simulated datasets? Sensitivity (%) is preferable to the number of DEGs in this case. The number of replicate is a discrete value

but the slope in this figure is smooth. What do the numbers in fold change (1.2 0.83-5, 0.2) mean?

4. Figure 3: Why were the result of 5x and 4x merged? What does each line indicate? # reads = 0 means nothing thus should be removed.

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

Partly

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?

Partly

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.

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, however I have significant reservations, as outlined above.

Author Response 10 Nov 2017

Binay Panda, Institute of Bioinformatics and Applied Biotechnology, Bangalore, India

We thank Dr. Daisuke Komura for his time and timely review. We have addressed all his queries below and have incorporated his suggestions in the revised version of the manuscript, which we believe is much improved now.

Response to Reviewers' comments

Major:

Reviewer's comments:

1) In general, some reads in RNA-seq analysis are removed due to their low sequencing quality or cannot be mapped probably due to library contamination of rRNA or other species' RNA. Does this software consider the effect? If not, it could underestimate the number of required reads.

Authors' response:

No, the software does not consider this effect. The simulated reads correspond to error-free/-corrected reads. The `add_error` and `add_platform_error` features are added in a recent release of Polyester, not available when we used Polyester.

Reviewer's comments:

2) How does RNAtor calculate the number of reads required for DEG detection? The authors claim that it is calculated based on the number of DEGs at its peak in both real and simulated data but the exact algorithm is not shown. Specifically, how were the peak values calculated and how were the peak values of real and simulated data merged?

Authors' response:

The detected DEGs by various tools were plotted as a function of simulated reads and replicates (Figure 2). The saturation point of the DEG curves, i.e. the peak was used to recommend a certain no of reads given the no of replicates, to detect optimal DEGs. Validation for this recommendation was obtained from extrapolation of the read numbers according to the *Saccharomyces* transcriptome size (Figure 3).

Minor:

Reviewer's comments:

3) Which does the term 'replicates' mean in this manuscript, technical replicate or biological replicate?

Authors' response:

We thank the reviewer to point this out. The replicates used were technical. We have made this change in the revised manuscript.

Reviewer's comments:

4) Implementation: "... workflow followed by differential expression analysis using five tools:..." I think four instead of five is correct because Kallisto does not use outputs of Tophat-Cufflinks pipeline (Supplementary Figure 1).

Authors' response:

Kallisto is used twice; first, with the genome-guided paradigm and second, with *de novo* assembly using Trinity. In the first scenario, the Tophat-Cufflinks alignments (.bam) were converted to reads (.fastq) to be used with Kallisto along with the 3MB transcriptome as the reference. In the second scenario, the *de novo* assembled transcriptome was used as the reference along with the simulated reads with Kallisto. This has been clarified in the revised manuscript.

Reviewer's comments:

5) Figure 2: How many DEGs were included in total in the simulated datasets? Sensitivity (%) is preferable to the number of DEGs in this case. The number of replicate is a discrete value but the slope in this figure is smooth. What do the numbers in fold change (1.2 0.83-5, 0.2)

mean? Figure 2: I think "fold change (1.2,0.83 - 5,0.2)" is confusing so it might be better to change it to another one such as (>1.2 or $<1/1.2$) to (>5 or $<1/5$)).

Authors' response:

The total no of simulated DEGs are 363. TP and FP curves are represented in Supplementary Figure S3. We have changed the representation for replicate numbers. The simulated fold changes were 1.2X to 5X, in both directions, i.e. $1.2, 1/1.2=0.83$ and $5, 1/5=0.2$. Fold change (1.2,0.83 - 5,0.2)" is changed to (>1.2 or $<1/1.2$) to (>5 or $<1/5$)).

Reviewer's comments:

6) Figure 3: Why were the result of 5x and 4x merged? What does each line indicate? # reads = 0 means nothing thus should be removed. Figure 3: What do the lines indicate in each subplot?

Authors' response:

The number of DEGs obtained separately at 4X and 5X were not sufficient to draw any conclusions, Hence they were merged as 4X+5X or $>4X$. We have removed the #reads=0 data point. We have updated the Figure 3 legend.

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research