SOFTWARE TOOL ARTICLE

**RNAtor:** an Android-based application for biologists to plan RNA sequencing experiments [version 1; referees: 1 approved with reservations]

Shruti Kane¹, Himanshu Garg¹, Neeraja M. Krishnan¹, Aditya Singh¹, Binay Panda ²¹,²

¹Ganit Labs, Bio-IT Centre, Institute of Bioinformatics and Applied Biotechnology, Bangalore, India
²Strand Life Sciences, Bangalore, India

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 **RNAtor**, a mobile application for Android platforms, to aid biologists in correctly designing their RNA-seq experiments. The recommendations from **RNAtor** are based on simulations and real data.

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.

How to cite this article: Kane S, Garg H, Krishnan NM et al. **RNAtor:** an Android-based application for biologists to plan RNA sequencing experiments [version 1; referees: 1 approved with reservations] *F1000Research* 2017, 6:997 (doi: 10.12688/f1000research.11982.1)

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

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.

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 RNAtor, an Android app with a user-friendly graphical user interface (GUI) that helps biologists design RNA-seq experiments. RNAtor 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. RNAtor’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 RNAtor (Figure 1). An RNAtor flowchart highlighting simulation conditions and analytical tools used is provided in Supplementary Figure S1.

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

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 RNAtor 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, RNAtor 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.
Table 1. **RNAtor** 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.

<table>
<thead>
<tr>
<th>Fold</th>
<th>2 replicates</th>
<th>3 replicates</th>
<th>4 replicates</th>
<th>5 replicates</th>
</tr>
</thead>
<tbody>
<tr>
<td>5x</td>
<td>6</td>
<td>2</td>
<td>1.5</td>
<td>1.5</td>
</tr>
<tr>
<td>4x</td>
<td>10</td>
<td>6</td>
<td>2</td>
<td>1.5</td>
</tr>
<tr>
<td>3x</td>
<td>10</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>2x</td>
<td>14</td>
<td>10</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>1.5x</td>
<td>30</td>
<td>20</td>
<td>20</td>
<td>14</td>
</tr>
</tbody>
</table>

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. RNAtor 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 RNAtor is available on Google Play Store.


License: RNAtor 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: RNAtor 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


Open Peer Review

Current Referee Status: ?

Version 1

Referee Report 06 July 2017
doi:10.5256/f1000research.12955.r24054

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 have read this submission. I 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.