FINAL SCRIPT: APPROVED FOR FILMING

Submission ID #: 69253
Scriptwriter Name: Sulakshana Karkala
Project Page Link: https://review.jove.com/account/file-uploader?src=21122213

Title: Rup (RNA-Seq Usability Assessment Pipeline) - Quality Control for Bulk RNA-Seq Experiments in Eukaryotes

Authors and Affiliations:
Oliver Rupp[1], Le-Han Roessner[2], Doudou Kong[2], Annette Becker[2]

[1] Bioinformatics and Systems Biology, Justus Liebig University
[2] Institute of Botany, Justus Liebig University

Corresponding Authors:
Annette Becker annette.becker@bot1.bio.uni-giessen.de
Email Addresses for All Authors:
Oliver Rupp: oliver.rupp@computational.bio.uni-giessen.de
Le-Han Roessner le-han.roessner@uni-giessen.de
Doudou Kong doudou.kong@outlook.com
Annette Becker annette.becker@bot1.bio.uni-giessen.de

Author Questionnaire

1. **Microscopy**: Does your protocol require the use of a dissecting or stereomicroscope for performing a complex dissection, microinjection technique, or something similar? **No**

2. **Software:** Does the part of your protocol being filmed include step-by-step descriptions of software usage? **Yes, all done**

3. **Filming location:** Will the filming need to take place in multiple locations? **No**

4. **Testimonials (optional):** Would you be open to filming two short testimonial statements **live during your JoVE shoot**? These will **not appear in your JoVE video** but may be used in JoVE's promotional materials. **No**

Current Protocol Length
Number of Steps: 11
Number of Shots: 23

Introduction

Videographer: Obtain headshots for all authors available at the filming location.

INTRODUCTION:

~~What is the scope of your research? What questions are you trying to answer?~~

1.1. **Oliver Rupp:** We develop innovative bioinformatics tools to simplify, automate, and integrate data analysis from diverse high-throughput experiments.

1.1.1. INTERVIEW: Named Talent says the statement above in an interview-style shot, looking slightly off-camera. *Suggested B.roll:3.4*

~~What technologies are currently used to advance research in your field?~~

1.2. **Oliver Rupp:** We use high-throughput sequencing, advanced bioinformatics software, and powerful computing infrastructures to enable systematic biological data analysis.

1.2.1. INTERVIEW: Named Talent says the statement above in an interview-style shot, looking slightly off-camera.

CONCLUSION:

~~What research gap are you addressing with your protocol?~~

1.3. **Oliver Rupp:** We address missing standardized quality control for RNA-seq, ensuring reliable data assessment before downstream gene expression analysis.

1.3.1. INTERVIEW: Named Talent says the statement above in an interview-style shot, looking slightly off-camera.

~~What advantage does your protocol offer compared to other techniques?~~

1.4. **Oliver Rupp:** Rup integrates multiple quality checks in one pipeline, offering accessible, automated, and reproducible RNA-seq assessment for biologists.

1.4.1. INTERVIEW: Named Talent says the statement above in an interview-style shot, looking slightly off-camera. *Suggested B.roll:4.2*

~~How will your findings advance research in your field?~~

1.5. **Oliver Rupp:** Our tool enables exploring how RNA-seq quality influences biological interpretation, paving the way for transparent and reproducible transcriptomics.

1.5.1. INTERVIEW: Named Talent says the statement above in an interview-style shot, looking slightly off-camera. *Suggested B.roll:4.3*

Videographer: Obtain headshots for all authors available at the filming location.

Protocol

2. Initial Setup and Configuration for Read Mapping Analysis

Demonstrator: Oliver Rupp

2.1. To begin, install all required R packages using the Bioconductor package manager **[1]**.

2.1.1. WIDE: Talent using the Bioconductor package manager on a desktop.

2.2. Create a source folder to organize the input files for the analysis **[1]**. Add the reference genome sequence in FASTA *(Fast-ah)* format as "reference/genome.fa" *(reference-genome-dot-f-a)* to this folder **[2]**.

2.2.1. SCREEN: 69253_screenshot_1.mp4. 00:00-00:09

2.2.2. SCREEN: 69253_screenshot_1.mp4. 00:10-00:27

2.3. Add the gene model annotation file named "reference/annotation.gtf" *(reference-annotation-dot-g-t-f)* to the same folder **[1]**. Optionally, include the rRNA *(R-R-N-A)* gene annotation as a GTF *(G-T-F)* file named "reference/rRNA.gtf" *(reference-R-N-A-Dot-G-T-F)* **[2]**.

2.3.1. SCREEN: 69253_screenshot_2.mp4 00:00-00:13.

2.3.2. SCREEN: 69253_screenshot_2.mp4 00:14-00:32

2.4. Place all sequencing reads as compressed fastq *(Fast-Q)* files into the folder named "reads" **[1]**. Ensure that each file follows the naming format **[2-TXT]**. Then set the analysis parameters according to the sequencing method used **[3]**.

2.4.1. SCREEN: 69253_screenshot_3.mp4. 00:00-00:05

2.4.2. SCREEN: 69253_screenshot_3.mp4. 00:06-00:20
TXT: File name format: <SAMPLENAME>_1.fastq.gz and <SAMPLENAME>_2.fastq.gz for the forward and reverse reads

2.4.3. SCREEN: 69253_screenshot_3.mp4. 00:21-00:27

3. Read Mapping Quality Assessment Using Rsubread

3.1. To map quality of the sequence, use the Rsubread *(R-S-U-Bread)* package **[1]** to build an index of the reference genome from the genome FASTA file **[2-TXT]**.

3.1.1. SCREEN: 69253_screenshot_4.mp4 00:00-00:06

3.1.2. SCREEN: 69253_screenshot_4.mp4. 00:10-00:17 **TXT: Perform this only once for each reference genome**

3.2. For each sample, use the **align()** *(Align)* function to iterate and align sequencing reads to the reference genome **[1]**. Store the resulting alignment files in the output folder in .bam *(Dot-Bam)* format **[2]**.

3.2.1. SCREEN: 69253_screenshot_5.mp4 00:00-00:09

3.2.2. SCREEN: 69253_screenshot_5.mp4 00:10-00:19

3.3. Now use the **featureCounts()** *(Feature-counts)* function to count reads mapped to each gene **[1].** The annotation files should be in the GTF format **[2]**. Ensure only reads with a single match to the genome are counted **[3]**.

3.3.1. SCREEN: 69253_screenshot_6.mp4. 00:00-00:14

3.3.2. SCREEN69253_screenshot_6.mp4. 00:15-00:25

3.3.3. SCREEN: 69253_screenshot_6.mp4. 00:29-00:40

3.4. Count the reads that map to rRNA genes by using the **featureCounts()** function with the rRNA gene GTF file **[1]**. Allow multimapped reads to be included in this count **[2]**.

3.4.1. SCREEN: 69253_screenshot_7.mp4. 00:00-00:16

3.4.2. SCREEN: 69253_screenshot_7.mp4. 00:17-00:26

3.5. Retrieve the read assignment statistics generated by the **featureCounts()** function for each sample **[1]**. These statistics include the number of reads categorized as assigned, unmapped, multimapped, and others **[2]**.

3.5.1. SCREEN: 69253_screenshot_8.mp4. 00:00-00:19

3.5.2. SCREEN: 69253_screenshot_8.mp4. 00:20-00:29

3.6. Collect the statistics for rRNA gene assignments separately **[1]**. Then generate bar plots visualizing the read mapping statistics from the previous steps **[2]**.

3.6.1. SCREEN: 69253_screenshot_9.mp4. 00:00-00:16

3.6.2. SCREEN: 69253_screenshot_9.mp4. 00:17-00:31

3.7. Group genes based on the number of reads assigned to them **[1].** Plot the classification results as a bar plot **[2]**.

3.7.1. SCREEN: 69253_screenshot_10.mp4. 00:00-00:12

3.7.2. SCREEN: 69253_screenshot_10.mp4. 00:16-00:26

Results

4. Results

4.1. Sample s1_r1 *(S-one-R-One)* showed a low number of reads both before and after trimming **[1]**.

4.1.1. LAB MEDIA: Figure 2. *Video editor: Highlight pink and blue bars for sample s1_r1 sequentially*

4.2. The trimmed read count of sample s1_r2 *(S-one-R-Two)* was visibly reduced compared to its raw read count **[1]**, indicating removal of low-quality reads during trimming **[2]**.

4.2.1. LAB MEDIA: Figure 2. *Video editor: Highlight the pink read bar for sample s1_r2.*

4.2.2. LAB MEDIA: Figure 2. *Video editor: Highlight the blue bar for sample s1_r2.*

4.3. Mapping identified problems in the read assignments **[1].** Sample s2_r3*(S-one-R-Three)* exhibited a high number of multi-mapped reads **[2]** and an elevated amount of ribosomal RNA reads **[3]**. A large fraction of reads in sample s2_r4 did not map to the reference genome suggesting contamination with sequences from a non-target organism**[4]**.

4.3.1. LAB MEDIA: Figure 3.

4.3.2. LAB MEDIA: Figure 3. *Video editor: Highlight the green portion of the stacked bar for sample s2_r3 in the "Genes" panel*

4.3.3. LAB MEDIA: Figure 3. *Video editor: Highlight the blue bar for sample s2_r3 in the "rRNA" panel.*

4.3.4. LAB MEDIA: Figure 3. *Video editor: Highlight the pink portion of the stacked bar for sample s2_r4 in the "genome" panel.*

4.4. Samples s2_r1 through s2_r4 showed fewer genes with more than 100 assigned reads **[1]**.

4.4.1. LAB MEDIA: Figure 4. *Video editor: Highlight the orange and red segments of the bars for samples s2_r1 through s2_r4.*

4.5. In the correlation heatmap, sample s2_r5 clustered with the replicates of sample s1 **[1]**, and sample s1_r5 clustered with the replicates of sample s2, indicating a likely replicate labeling error **[2].**

4.5.1. LAB MEDIA: Figure 5. *Video editor: Highlight the red box of s2_r5 row*

4.5.2. LAB MEDIA: Figure 5. *Video editor: Highlight the red box of of s1_r5 within the* *cluster of s2 replicates on the heatmap.*

Pronunciation Guide:

1. RNA-seq
   Pronunciation link: No confirmed link found
   IPA: /ˌɑːrˌɛnˈeɪ sɪk/
   Phonetic Spelling: ar-eh-nay-seek
2. Eukaryotes
   Pronunciation link: https://www.merriam-webster.com/dictionary/eukaryote
   IPA: /juːˈkær.i.oʊts/
   Phonetic Spelling: yoo-kair-ee-oats
3. Bioinformatics
   Pronunciation link: https://www.merriam-webster.com/dictionary/bioinformatics
   IPA: /ˌbaɪ.oʊˌɪn.fərˈmæt.ɪks/
   Phonetic Spelling: bye-oh-in-fer-mat-iks
4. Transcriptomics
   Pronunciation link: No confirmed link found
   IPA: /ˌtrænˌskrɪpˈtoʊ.mɪks/
   Phonetic Spelling: tran-skript-oh-miks
5. FASTA
   Pronunciation link: No confirmed link found
   IPA: /ˈfæs.tə/
   Phonetic Spelling: fas-tuh
6. GTF
   Pronunciation link: No confirmed link found
   IPA: /ˌdʒiː.tiːˈɛf/
   Phonetic Spelling: jee-tee-ef
7. FASTQ
   Pronunciation link: No confirmed link found
   IPA: /ˈfæst.kjuː/
   Phonetic Spelling: fast-kyoo
8. rRNA
   Pronunciation link: No confirmed link found
   IPA: /ˌɑːrˌɑːrɛnˈeɪ/
   Phonetic Spelling: ar-ar-en-ay
9. Rsubread
   Pronunciation link: No confirmed link found
   IPA: /ɑːrˈsʌbˌriːd/
   Phonetic Spelling: ar-sub-reed
10. featureCounts
    Pronunciation link: No confirmed link found
    IPA: /ˈfiː.tʃərˌkaʊnts/
    Phonetic Spelling: fee-chur-kownts
11. Multimapped
    Pronunciation link: No confirmed link found

IPA: /ˌmʌl.tiˈmæpt/
Phonetic Spelling: mul-tee-mapt

12. Annotation
    Pronunciation link: https://www.merriam-webster.com/dictionary/annotation
    IPA: /ˌæn.əˈteɪ.ʃən/
    Phonetic Spelling: an-uh-tay-shun
13. Ribosomal
    Pronunciation link: https://www.merriam-webster.com/dictionary/ribosomal
    IPA: /ˌraɪ.bəˈsoʊ.məl/
    Phonetic Spelling: rye-buh-so-mul