AUTHOR PROVIDED FOOTAGE

Submission ID #: 61633
Scriptwriter Name: Anastasia Gomez
Project Page Link: https://www.jove.com/account/file-uploader?src=18798113

Title: A bioinformatics pipeline for investigating molecular evolution and gene expression using RNA-seq

Authors and Affiliations:

Aide Macias-Muñoz[1*], Ali Mortazavi[1,*]

[1] Department of Developmental and Cell Biology, University of California, Irvine, CA, U.S.A.

Corresponding Authors:
ali.mortazavi@uci.edu
amaciasm@uci.edu

Author Questionnaire

This is an Author Provided Footage Video

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

3. Interview statements: Considering the COVID-19-imposed mask-wearing and social distancing recommendations, which interview statement filming option is the most appropriate for your group? **Please select one**.

☒ Interviewees self-record interview statements. JoVE can provide support for this option.

Current Length

Total Number of Steps: 19

Introduction

1. Introductory Interview Statements

REQUIRED:

1.1. **Ali Mortazavi:** This protocol outlines bioinformatic steps for investigating the molecular evolution and expression of candidate genes.

1.1.1. INTERVIEW: Named talent says the statement above in an interview-style shot, looking slightly off-camera. NOTE: All interview statements uploaded to AWS project page: https://www.jove.com/account/file-uploader?src=18798113

1.2. **Aide Macias-Muñoz:** Here, we provide thorough instructions for using multiple programs so that someone with minimal bioinformatics experience could follow this protocol.

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

OPTIONAL:

1.3. **Aide Macias-Muñoz:** This pipeline can be applied to any organism and any gene family of interest.

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

1.4. **Aide Macias-Muñoz:** One common issue in bioinformatics is shell scripts failing. When attempting this protocol, make sure to read manuals carefully, check error files, and use the most up-to-date programs if compatible.

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

Ethics Title Card

1.5. Procedures involving animal subjects have been approved by the Institutional Animal Care and Use Committee (IACUC) at UC Irvine.

Protocol

2. Obtain RNA-seq Reads and Reference Assembly

2.1. To begin, login to the computer cluster account on a terminal or PuTTY *(pronounce like silly putty)* application window **[1]**. On the terminal, download SRA Toolkit version 2.8.1 using **wget** *(pronounce 'W-get')*, then finish installing the program **[2]**.

2.1.1. SCREEN: 18798113_screenshot_1.mov. *Video Editor: Emphasize the command being typed in.*

2.1.2. SCREEN: 18798113_screenshot_2.mov. *Video Editor: Emphasize the command being typed in.*

2.2. Search NCBI for the SRA accession number for the desired samples **[1-TXT]**, then obtain the RNA-seq data in the terminal window **[2]**. Obtain two fastq *(pronounce 'fast-Q')* files for paired end files type **[3-TXT]**.

2.2.1. SCREEN: 18798113_screenshot_3.mov. **TEXT: Format: SRRXXXXXX**

2.2.2. SCREEN: 18798113_screenshot_4.mov. *Video Editor: Emphasize the command being typed in.*

2.2.3. SCREEN: 18798113_screenshot_5.mov. **TEXT: SRRXXXXXX_1.FASTQ and SRRXXXXXX_2.FASTQ** *Video Editor: Emphasize the command being typed in.*

2.3. To obtain a reference assembly, type **wget** in the terminal window and paste the link address. If available, also copy the GTF file and protein FASTA file for the reference genome **[1]**.

2.3.1. SCREEN: 18798113_screenshot_6.mov.

2.4. Index the genome **[1]**, then map reads and calculate expression for each sample **[2]**. Rename the results file to something descriptive **[3]** and generate a matrix of all counts **[4]**.

2.4.1. SCREEN: 18798113_screenshot_7.mov. *Video Editor: Emphasize the command being typed in.*

2.4.2. SCREEN: 18798113_screenshot_8.mov. *Video Editor: Emphasize the command being typed in.*

2.4.3. SCREEN: 18798113_screenshot_9.mov. *Video Editor: Emphasize the command being typed in.*

2.4.4. SCREEN: 18798113_screenshot_10.mov. *Video Editor: Emphasize the command being typed in.*

3. Identify Genes of Interest

3.1. Open an internet browser window and go to NCBI GenBank **[1-TXT]**. In the search bar, type the name of the gene of interest and the name of closely related species which have been sequenced. On the left of the search bar select protein, then click search **[2]**.

3.1.1. SCREEN: 18798113_screenshot_11.mov. 0:00 – 0:05. **TEXT: https://www.ncbi.nlm.nih.gov/genbank/**

3.1.2. SCREEN: 18798113_screenshot_11.mov. 0:05 – 0:14.

3.2. Extract the sequences by clicking **Send to** and then select **File**. Under Format, select FASTA then click **Create File [1]**.

3.2.1. SCREEN: 18798113_screenshot_11.mov. 0:15 – end.

3.3. Move FASTA file of homologs to the computer cluster using a local terminal window or FileZilla **[1]**.

3.3.1. SCREEN: 18798113_screenshot_12.mov. *Video Editor: Emphasize the command being typed in.*

3.4. Next, search for candidate genes using BLAST+. On the computer cluster, make a BLAST database from the genome or transcriptome translated protein FASTA **[1]**.

3.4.1. SCREEN: 18798113_screenshot_13.mov. *Video Editor: Emphasize the command being typed in.*

3.5. BLAST the homologous gene sequences from NCBI to the database of the species of interest, then view the output file using the command **more**. Copy unique gene IDs from the species of interest to a new text file **[1]**. Extract the sequences of candidate genes **[2]**.

3.5.1. SCREEN: 18798113_screenshot_14.mov.

3.5.2. SCREEN: 18798113_screenshot_15.mov.

3.6. To confirm gene annotation using reciprocal BLAST, go to the BLAST Local Alignment Search Tool **[1-TXT]**, select blastx *(pronounce 'blast-X')*, then paste the candidate sequences, select the Non-redundant protein sequence database and click **BLAST [2]**.

3.6.1. SCREEN: 18798113_screenshot_16.mov. 0:01. **TEXT: https://blast.ncbi.nlm.nih.gov/Blast.cgi**

3.6.2. SCREEN: 18798113_screenshot_16.mov. 0:02 – end.

4. Phylogenetic Trees

4.1. Open metrics prefix mega, click on **Align**, then **Edit-Build Alignment**, select **Create a new alignment** and click **OK**. Select **Protein**. When the alignment window opens, click on **Edit**. Click **Insert sequences from file** and select the FASTA with protein sequences of candidate genes and probable homologs **[1]**.

4.1.1. SCREEN: 18798113_screenshot_17.mov. 0:00 – 0:19.

4.2. Select all sequences. Find the arm symbol and hover over it, it should say Align sequences using MUSCLE algorithm. Click on the arm symbol and then click **Align Protein** to align the sequences. Edit parameters or click **OK** to use default parameters **[1]**.

4.2.1. SCREEN: 18798113_screenshot_17.mov. 0:20 – end.

Results

5. **Results: Opsin genes in *Hydra vulgaris***

5.1. This protocol was applied to tissues of *Hydra vulgaris*, which is a fresh-water invertebrate that belongs to the phylum *Cnidaria*. Opsin genes were investigated to gain insight into the evolution of eyes and light detection in animals **[1]**.

5.1.1. LAB MEDIA: Figure 1.

5.2. Sequences for opsin-related genes of *H. vulgaris* and other species were extracted into a fasta file from the NCBI GenBank. The opsin genes were aligned in metrics prefix mega, making it possible to identify *Hydra* opsins that were missing a conserved lysine amino acid necessary to bind a light sensitive molecule **[1]**.

5.2.1. LAB MEDIA: Figure 2.

5.3. A maximum-likelihood tree was generated using opsin sequences from *Hydra vulgaris* and other species **[1-TXT]**.

5.3.1. LAB MEDIA: Figure 3. *Video Editor: Display the other species names as a text overlay here.* **TEXT: *Podocoryna carnea*, *Cladonema radiatum*, *Tripedelia cystophora*, *Nematostella vectensis, Mnemiopsis leidyi, Trichoplax adhaerens*, *Drosophila melanogaster* and *Homo sapiens***

5.4. The phylogeny suggests opsin genes are evolving by lineage-specific duplications in cnidarians and potentially by tandem duplication in *H. vulgaris* **[1]**.

5.4.1. LAB MEDIA: Figure 3.

5.5. Next, a differential expression analysis was performed in edgeR to investigate absolute expression of opsin genes. To determine whether one or more opsins are upregulated in the hypostome, or head, pair-wise comparisons of hypostome versus the body column, budding zone, foot, and tentacles were performed **[1]**.

5.5.1. LAB MEDIA: Table 1.

5.6. It was found that 1,774 transcripts were differentially expressed between the hypostome and body column. The genes that were upregulated across multiple comparisons were determined and a functional enrichment in Blast2GO was performed **[1]**.

5.6.1. LAB MEDIA: Table 1.

5.7. Finally, the absolute expression of opsin genes was investigated in different tissues **[1]**, during different stages of budding **[2]**, and during different time points of regeneration **[3]**.

5.7.1. LAB MEDIA: Figure 4. *Video Editor: Emphasize A.*

5.7.2. LAB MEDIA: Figure 4. *Video Editor: Emphasize B.*

5.7.3. LAB MEDIA: Figure 4. *Video Editor: Emphasize C.*

Conclusion

6. Conclusion Interview Statements

NOTE: All interview statements uploaded to AWS project page: https://www.jove.com/account/file-uploader?src=18798113

6.1. **Aide Macias-Muñoz:** Visual inspection of an alignment and tree will confirm whether the candidate genes belong to the family of interest. Genes that are too different in sequence and group outside probably belong to a different gene family.

6.1.1. INTERVIEW: Named talent says the statement above in an interview-style shot, looking slightly off-camera. *Suggested B-roll: Shot that corresponds with 5.1.1 in the text manuscript.*

6.2. **Aide Macias-Muñoz:** Results from this protocol can be used to identify genes important to organisms, which can then be validated for function in future experiments.

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

6.3. **Aide Macias-Muñoz:** After exploring Hydra opsin expression, we are now using similar techniques to investigate related genes across species in order to identify differences and similarities in function.

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