RNA-Seq Demo on Galaxy

Tom Doak Sheri Sanders Bhavya NP
Carrie Ganote
National Center for Genome Analysis Support
Our RNA-Seq Demo Data

We will be assembling two conditions of Yeast - diauxic shift and heat shock. We’ll refer to these as ds and hs for the class.

Schizosaccharomyces pombe (fission yeast). The data are paired-end 76bp RNA-Seq reads.

I’m following the tutorial from Trinity’s github page:

https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/wiki/Trinity-De-novo-Transcriptome-Assembly-Workshop
Let’s get some sequence data

Galaxy allows users to publish their data to share with each other.

Let’s start with “Shared Data” at the top. Then select Data Libraries from the menu.
Let’s get some sequence data

Choose Workshop Data.

Then Bioinformatics2Go

National Center for Genome Analysis Support: http://ncgas.org
Let’s get some sequence data

Check the box

Then click “to History”

National Center for Genome Analysis Support: http://ncgas.org
Let’s get some sequence data

You may choose to add the data to an existing history or create a new one.

National Center for Genome Analysis Support: http://ncgas.org
Let's get some sequence data

Data set is imported!
Click on the green button
to go to it..

If you missed the green button (it disappears quickly!),
you can always get back to the home page by clicking “Analyze Data”.

National Center for Genome Analysis Support: http://ncgas.org
Who doesn’t see this?

Thank you for choosing Galaxy!
Slides for our July Galaxy talks (part of the Bioinformatics Clinic) are available:
- A Short Demo on RNA-Seq using the Tuxedo Suite
- A Short Demo on RNA-Seq using Trinity
- Guide to Workflows for Automating Galaxy
- Moving Large Data onto Galaxy
- Galaxy for Data Provenance

National Center for Genome Analysis Support: http://ncgas.org
Step 1: Assess the Quality of Inputs

We will first get an idea of the quality of our input data sets.

The FastQC tool will produce graphical output that makes it easy to gauge the characteristics of the data – quality, patterns, biases, gc content etc.

Quality Control
- Trim Galore! Quality and adapter trimmer of reads
- Trimmomatic flexible read trimming tool for Illumina NGS data
- Trim sequences
- Cutadapt Remove adapter sequences from Fastq/Fasta
- Compute quality statistics
- Draw nucleotides distribution chart
- Filter Fastq reads by quality score and length
- Fastq Quality Trimmer by sliding window
- Fastq Groomer convert between various FASTQ quality formats
- FastQC Read Quality reports
- Bar chart for multiple columns
- Boxplot of quality statistics

Pro tip: Use the search bar to find tools
Step 1: Assess the Quality of Inputs

Choose any left or right reads file and run it. Compare your results with your neighbors’.

Leave these as defaults.
The input data usually declines in quality as the reads progress.

The quality score is assigned by the sequencing machine as it reads each base. It is a rough estimate of how ambiguous the signal is.

Sequence: ATGCATG
Quality Score: 39 38 23 19 3 3

National Center for Genome Analysis Support: http://ncgas.org
Step 2: Trim Input Sequences

We’ve determined that the input data sets need some work before they are used in downstream processes.

We’ll use Trimmomatic to trim reads based on quality score.

Run it on each Pair.

Quality Control
- Trim Galore! Quality and adapter trimmer of reads
- Trimmomatic flexible read trimming tool for Illumina NGS data
- Trim sequences
- Cutadapt Remove adapter sequences from Fastq/Fasta
- Compute quality statistics
- Draw nucleotides distribution chart
- Filter Fastq reads by quality score and length
- Fastq Quality Trimmer by sliding window
- Fastq Groomer convert between various FASTQ quality formats
- FastQC Read Quality reports
- Bar chart for multiple columns
- Boxplot of quality statistics

National Center for Genome Analysis Support: http://ncgas.org
Step 2: Trim Input Sequences

Select Paired-end reads (two files).

Left = R1, Right = R2.

2: Trimmomatic Operation

Select Trimmomatic operation to perform

Drop reads below a specified length (MINLEN)

Minimum length of reads to be kept

26

Insert an Operation to set the Minimum Length to 26.

National Center for Genome Analysis Support: http://ncgas.org
Step 2: Trim Input Sequences

Finally, set the average quality score to 25.

Your changed settings should look like the ones to the right.
Pro Tip: Rerunning Jobs

If you have to repeat a task but with different inputs, use the rerun feature.

Click on one of the outputs from the Galaxy run and look for the two arrows.

This allows you to easily check what parameters you used before.

National Center for Genome Analysis Support: http://ncgas.org
Pro Tip: Tidying History

You can hide or delete datasets that you know you don’t want to use in the future.

This does not delete anything from disk unless you take an extra step. You can undelete if you need to.

Take out the (R1 unpaired) and (R2 unpaired) results using the X.
Step 3: Rinse, Repeat

Now that the files are trimmed, we will re-assess their quality. If necessary, keep trimming away until you are satisfied with the input files.

Run FastQC again on the newly trimmed version of the file you ran before.
Step 3: Rinse, Repeat

Pictured are some left and right reads after trimming is complete. These will do!
Next we will put the reads together to create a complete picture of the actively transcribed genes of the sample organism.

Trinity is a de novo assembler that gives good results for RNA-seq. We will use it to assemble our reads.
Step 4: Assembly

Use ctrl or command + click to select multiple datasets at once. You may also shift+click to select blocks.

Select the two left files and the two right files for their respective boxes.

No other options need to be set!
Step 5: Assessing Quality of Assembly

Important statistics for assembly quality:
Contig Length Distribution

Assemblies will typically produce a number of complete contigs representing whole transcripts, and a large number of partial transcripts. This biases the average contig length toward the low end. The N50 is a measure weighted by total sequence length in the assembly.

Total Sequence: 100bp

Contigs, sorted by length

| 14 | 12 | 11 | 10 | 9 | 8 | 7 | 5 | 4 | 3 | 2 | 2 | 11111111 |

N50 = 9

Average Contig Length = 4.5

National Center for Genome Analysis Support: http://ncgas.org
Step 5: Assessing Quality of Assembly

Getting these stats in Galaxy:

Run assemblystats to get a summary and histograms of your contig length distribution.

National Center for Genome Analysis Support: [http://ncgas.org](http://ncgas.org)
Step 5: Assessing Quality of Assembly

Type of read
- Isotig (if from transcriptomic assembly)

Is this from an genomic (contig) or transcriptomic assembly (isotig) or are these raw reads (read)

Output histogram with bin sizes=1
- Yes
- No

Use this to specify whether or not bin sizes of 1 should be used when plotting histograms

Source file in FASTA format

Return all output files
- Yes
- No

If checked, all output files will be displayed. If not checked, only the file 'Assembly Statistics' will be provided.

Choose Isotig (since this is RNA) and return all output files. Make sure you are looking at the Trinity assembled transcripts

National Center for Genome Analysis Support: [http://ncgas.org](http://ncgas.org)
Step 8: Differential Expression

We threw everything into the Trinity assembly, but now we need to compare each Condition to that assembly using RSEM.

You will want to run RSEM twice – once for:
- ds left, ds right
And for:
- hs left, hs right
Step 8: Differential Expression

Here's an example for the DS dataset.

Rerun the job and change the two reads files to run the HS set as well.
Pro Tip: Renaming History Items

Let’s rename the RSEM results to RSEM HS isoform counts, HS gene counts, DS isoform counts, and DS gene counts.

Use the pencil next to the name of the history item, change the name, then click on Save attributes.
Step 8: Differential Expression

Next, we want to put these two together so we can look at the counts side-by-side.

The output is a ‘matrix’, just a table of counts.

de-novo RNAseq
- Generate gene to transcript map for Trinity assembly
- DESeq2 Determines differentially expressed features from count tables
- STAR-Fusion detect fusion genes in RNA-Seq data
- htsseg-count - Count aligned reads in a BAM file that overlap features in a GFF file
- RSEM abundance estimation run RSEM to estimate transcript abundances
- abundance estimation to matrix Join RSEM estimates from multiple samples into a single matrix
- EdgeR differentialExpression Identify Differentially Expressed Transcripts Using EdgeR
- Analyze Differential Expression Analyze differential expression
Step 8: Differential Expression

Add two estimates and use the Gene counts from each of your RSEM runs. Make sure the labels make sense and don’t use zany characters.
Step 8: Differential Expression

Now we’ll use EdgeR to see if there is a significant difference between the counts in one condition vs. the other, for each gene.
Step 8: Differential Expression

Use the Counts matrix and the Trinity assembled transcripts.

EdgeR_differentialExpression Identify Differentially Expressed Transcripts Using EdgeR (Galaxy Version 0.0.1)

Matrix of RNA-Seq fragment counts for transcripts per condition

20: abundance_estimation_to_matrix on data 19 and data 17: Counts Matrix

Transcripts fasta file corresponding to matrix

9: Trinity on data 2, data 6, and others: Assembled Transcripts

dispersion value

0.1

Dispersion value to be used in the negative binomial

Execute

National Center for Genome Analysis Support: http://ncgas.org
Step 8: Differential Expression

The last step is to visualize the results from the statistical analysis.
Step 8: Differential Expression

Use the EdgeR results and the TMM matrix to do this.

![Differential Expression Analysis Form](image)
Step 8: Differential Expression

You should get some kind of pretty heatmap – the hard part now is to interpret your results =)
Step ..?

RNA-Seq is a very versatile technology. You can use the data for:

- Gene discovery based on transcripts
- Genome evidence – introns, exons, junction
- Gene expression patterns in multiple samples
- SNP calling/other variants
- Protein divergence between samples

We have gotten to the assembly step, but there is a lot to learn about the data now that it is put together. A foundation in the use of Galaxy coupled with Indiana University resources will enable you to reach these goals.
Fin

Thanks for watching!
Questions and comments:
Email help@ncgas.org