

RNA-Seq Demo on Galaxy

Tom Doak Sheri Sanders Bhavya NP

Carrie Ganote

National Center for Genome Analysis Support

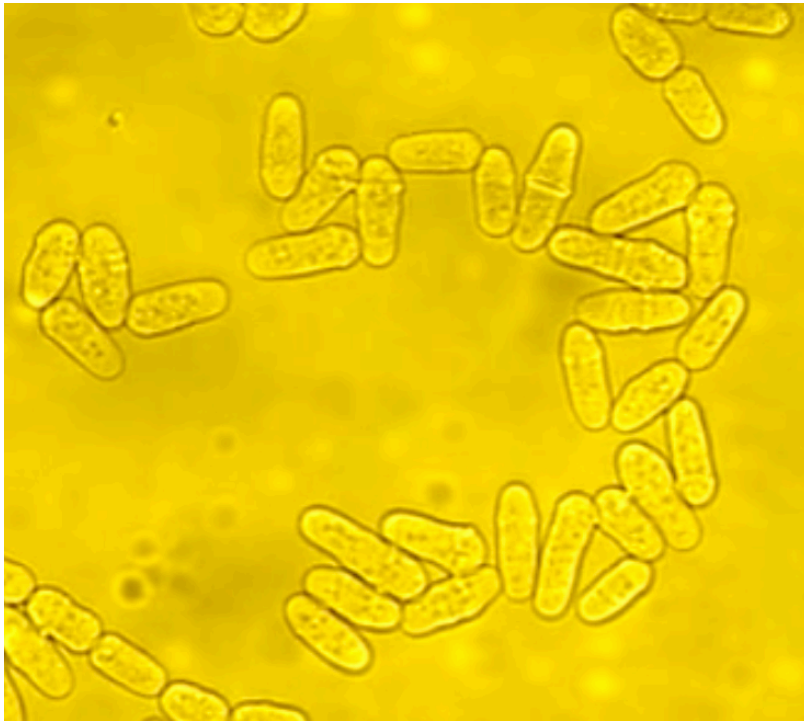


INDIANA UNIVERSITY



INDIANA UNIVERSITY

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:

Fission Yeast courtesy of Dr. Takeshi Hayashi

https://github.com/trinityrnaseq/RNASeq_Trinity_Tuxedo_Workshop/wiki/Trinity-De-novo-Transcriptome-Assembly-Workshop

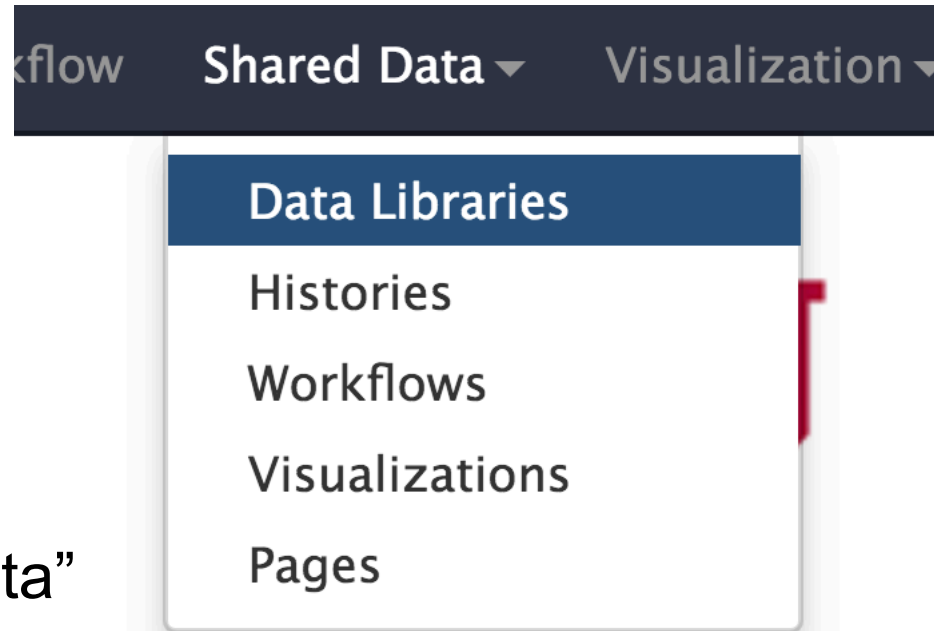
National Center for Genome Analysis Support: <http://ncgas.org>



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.





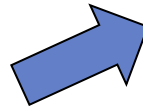
Let's get some sequence data

The screenshot shows the Galaxy web interface. At the top is a navigation bar with 'Galaxy' and tabs for 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', and 'Admin'. Below this is the 'Data Libraries' section, which includes a search bar and a table of libraries. The table has two columns: 'Data library name' and 'Data library description'. The 'Workshop Data' link is circled in yellow.

Data library name ↓	Data library description
User Import Library	For moving large datasets into Galaxy
Workshop Data	Learning sets of RNA-Seq data

Choose Workshop Data.

Then Bioinformatics2Go



This screenshot shows a dropdown menu for selecting a data library. It has a header with a checkbox and the text 'name ↓'. Below the header is a list of libraries, each with a folder icon, a checkbox, and the library name. The 'Bioinformatics2Go' option is highlighted.

<input type="checkbox"/>	name ↓
<input type="checkbox"/>	..
<input type="checkbox"/>	Bioinformatics2Go
<input type="checkbox"/>	Galaxy Workshop 2016
<input type="checkbox"/>	Galaxy Workshop 2018



INDIANA UNIVERSITY

Let's get some sequence data

Check
the box

Then click “to History”

DATA LIBRARIES << 0 1 2 >> showing 5 of 5 items ☐ include deleted + + to History Download x Delete i Details ? Help

Libraries / Galaxy Workshop 2018 / Galaxy Workshop 2018

<input checked="" type="checkbox"/> name ↓	description	data type	size	time updated (UTC)
<input checked="" type="checkbox"/> genome.fa		fasta	448.0 KB	2018-02-22 04:40 PM
<input checked="" type="checkbox"/> Sp_ds.left.fq		fastqsanger	17.0 MB	2018-02-22 04:37 PM
<input checked="" type="checkbox"/> Sp_ds.right.fq		fastqsanger	17.0 MB	2018-02-22 04:37 PM
<input checked="" type="checkbox"/> Sp_hs.left.fq		fastqsanger	18.0 MB	2018-02-22 04:37 PM
<input checked="" type="checkbox"/> Sp_hs.right.fq		fastqsanger	18.0 MB	2018-02-22 04:37 PM

<< 0 1 2 >> showing 5 of 5 items

National Center for Genome Analysis Support: <http://ncgas.org>



Let's get some sequence data

Import into History

Select history:

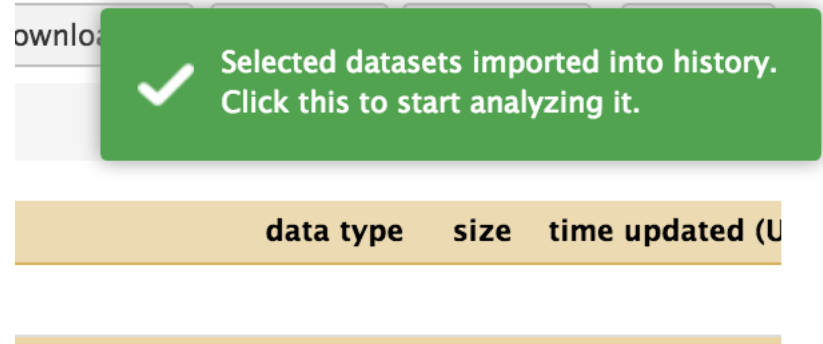
or create new:

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

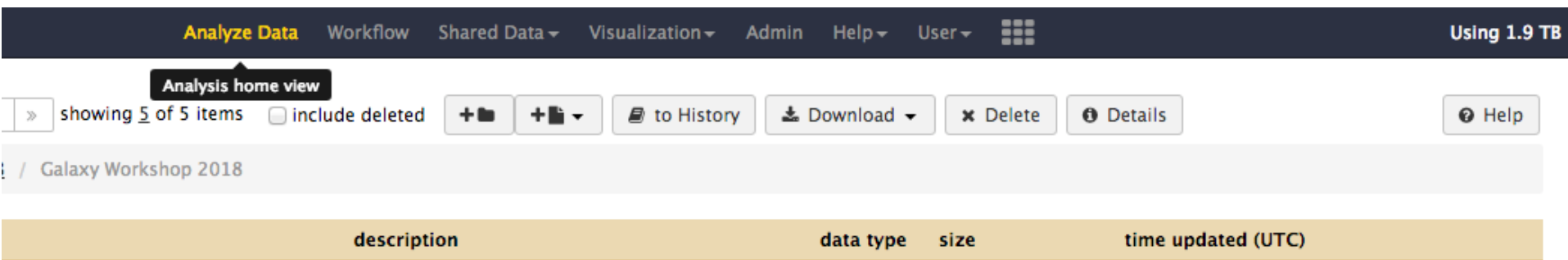


Let's get some sequence data

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”.





INDIANA UNIVERSITY

Who doesn't see this?

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User Using 1.9 TB

Tools

search tools

[Get Data](#)
[Send Data](#)
[Collection Operations](#)
[Text Manipulation](#)
[Fastq manipulation tools](#)
[Convert Formats](#)
[Quality Control](#)
[DNA Assembly](#)
[RNA Assembly](#)
[Assembly Statistics](#)
[Picard](#)
[Bedtools](#)
[Samtools](#)
[Alignment tools](#)
[Tuxedo Suite](#)
[de-novo RNAseq](#)
[Run BLAST+](#)

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT
INDIANA UNIVERSITY

✓ Welcome to the Galaxy Instance at Indiana University

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](#)

History

search datasets

Unnamed history
5 shown
448 KB

5: Sp hs.right.fg
4: Sp hs.left.fg
3: Sp ds.right.fg
2: Sp ds.left.fg
1: genome.fa

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

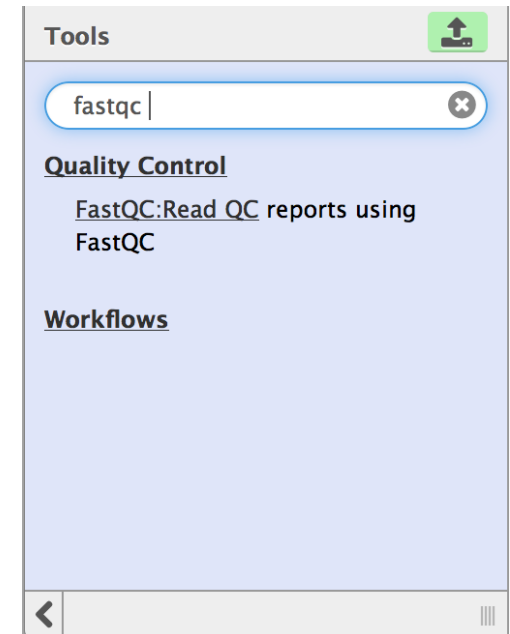
Fastq Groomer convert between various FASTQ quality formats

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.

FastQC Read Quality reports (Galaxy Version 0.70) Options

Short read data from your current history

5: Sp_hs.right.fq

Contaminant list

Nothing selected

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT
Primer CAAGCAGAAGACGGCATACGA

Submodule and Limit specifying file

Nothing selected

a file that specifies which submodules are to be executed (default=all) and also specifies the
thresholds for the each submodules warning parameter

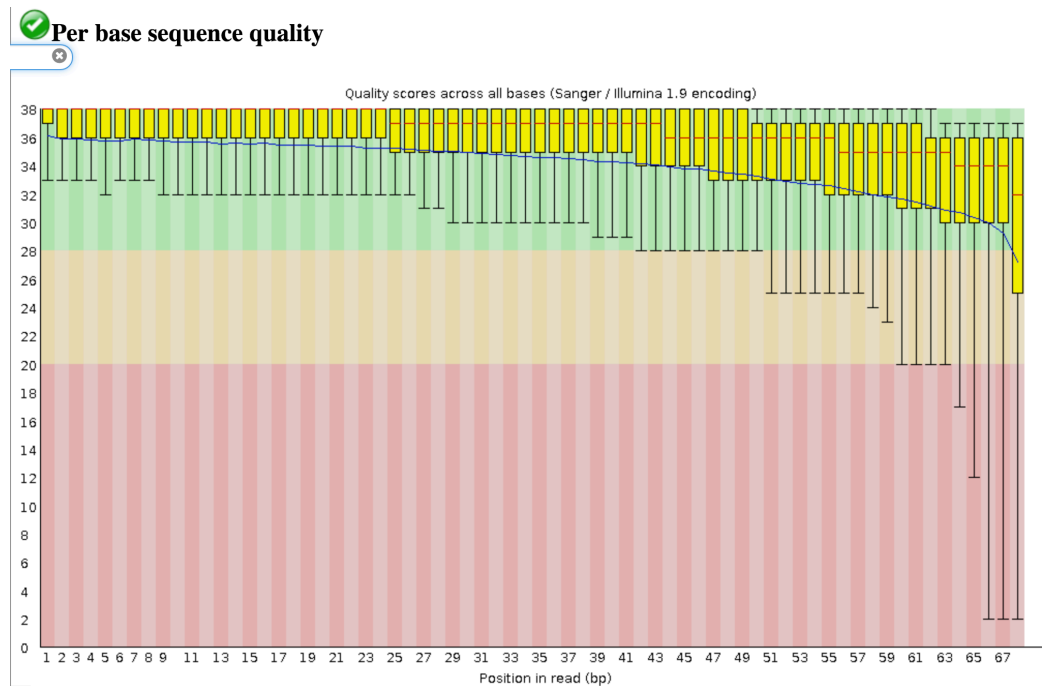
✓ Execute



Step 1: Assess the Quality of Inputs

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



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



Step 2: Trim Input Sequences

Select Paired-end reads (two files).

Left = R1, Right = R2.

Single-end or paired-end reads?

Paired-end (two separate input files)

|

Single-end

Paired-end (two separate input files)

Paired-end (as collection)

2: Trimmomatic Operation

Select Trimmomatic operation to perform

Drop reads below a specified length (MINLEN)

Minimum length of reads to be kept

26

+ Insert Trimmomatic Operation

Insert an Operation to set the Minimum Length to 26.



Step 2: Trim Input Sequences

Finally, set the average quality score to 25.

Your changed settings should look like the ones to the right.

Trimmomatic flexible read trimming tool for Illumina NGS data (Galaxy Version 0.36.4) Options

Single-end or paired-end reads?

Paired-end (two separate input files)

Input FASTQ file (R1/first of pair)

2: Sp_ds.left.fq

Input FASTQ file (R2/second of pair)

3: Sp_ds.right.fq

Trimmomatic Operation

1: Trimmomatic Operation

Select Trimmomatic operation to perform

Sliding window trimming (SLIDINGWINDOW)

Number of bases to average across

4

Average quality required

25

2: Trimmomatic Operation

Select Trimmomatic operation to perform

Drop reads below a specified length (MINLEN)

Minimum length of reads to be kept

26

+ Insert Trimmomatic Operation






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.

6: Trimmomatic on Sp...hs.left.fq (R1 paired)   

16.0 MB
format: **fastqsanger**, database: ?

TrimmomaticPE: Started with arguments:
-threads 4 -phred33
fastq_r1.fastqsanger
fastq_r2.fastqsanger
fastq_out_r1_paired.fastqsanger
fastq_out_r1_unpaired.fastqsanger
fastq_out_r2_paired.fastqsanger
fastq_out_r2_unpaired.fastqsanger
SLIDINGWINDOW:4:20 M

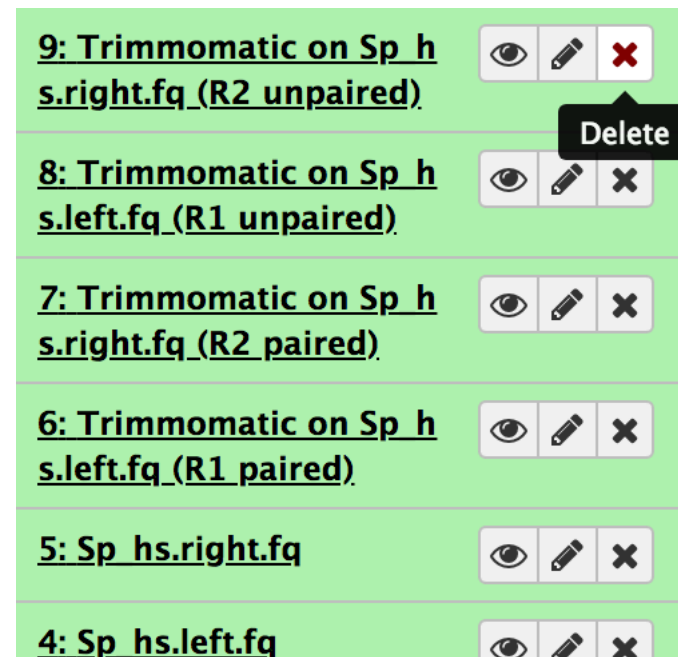
Run this job again :100:10004:15548/1
CGCCACGAGGTGCAGGGGGGATACCGGAAAGCTCA/



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.



INDIANA UNIVERSITY

Step 3: Rinse, Repeat

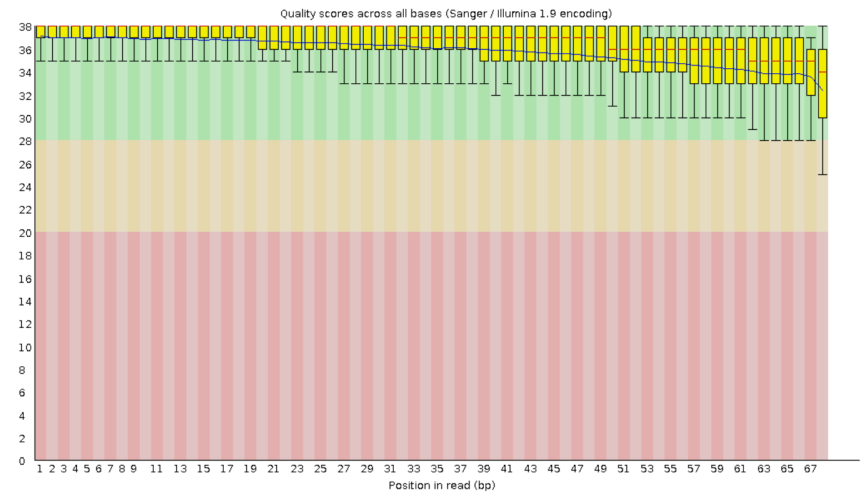
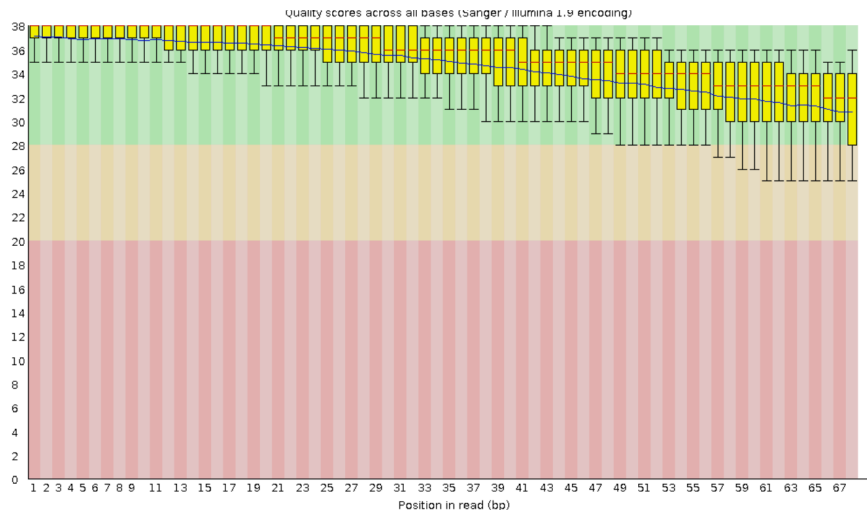
Now that the files are trimmed, we will re-assess their quality. If necessary, keep trimming away until you are satisfied with the input files.

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!





Step 4: Assembly

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.

The screenshot displays the Galaxy web interface for the Trinity de novo assembly tool. On the left, the 'Tools' sidebar lists various categories, with 'RNA Assembly' and 'Trinity de novo assembly of RNA-Seq data' highlighted by a yellow circle. The main panel shows the configuration for 'Trinity de novo assembly of RNA-Seq data (Galaxy Version 2.4.0.0)'. The 'Paired or Single-end data?' dropdown is set to 'Paired'. Under 'Left/Forward strand reads', the input is '25: genome.fa' with a list of transcripts: '13: Trinity on data 10, data 6, and others: Assembled Transcripts', '10: Trimmed Sp_hs.right', '9: Trimmed on Sp_hs.left', and '6: Trimmed Sp_ds.right'. The 'Right/Reverse strand reads' section is identical. The 'Strand specific data' section has 'Yes' and 'No' buttons. The 'Jaccard Clip options' section has 'Yes' and 'No' buttons, with a note 'set if you expect high gene density with UTR overlap (--jaccard_clip)'. The 'Run in silico normalization of reads' section has 'Yes' and 'No' buttons, with a note 'Defaults to max. read coverage of 50. (--normalize_reads)'. At the bottom, there is an 'Additional Options' section and an 'Execute' button.



Step 4: Assembly

Left/Forward strand reads



25: genome.fa
13: Trinity on data 10, data 6, and others: Assembled Transcripts
10: Trimmed Sp_hs.right
9: Trimmed on Sp_hs.left
6: Trimmed Sp_ds.right
5: Trimmed Sp_ds.left
4: Sp_hs.right.fq
3: Sp_hs.left.fq
2: Sp_ds.right.fq
1: Sp_ds.left.fq

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

(--left)

Right/Reverse strand reads



25: genome.fa
13: Trinity on data 10, data 6, and others: Assembled Transcripts
10: Trimmed Sp_hs.right
9: Trimmed on Sp_hs.left
6: Trimmed Sp_ds.right
5: Trimmed Sp_ds.left
4: Sp_hs.right.fq
3: Sp_hs.left.fq
2: Sp_ds.right.fq
1: Sp_ds.left.fq

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

No other options need to be set!

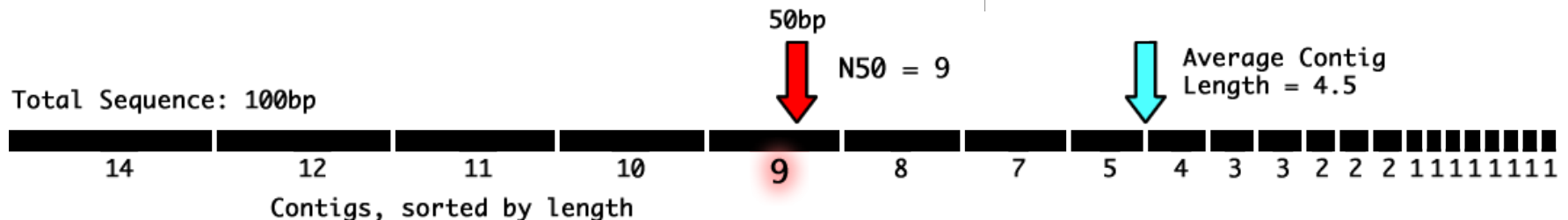
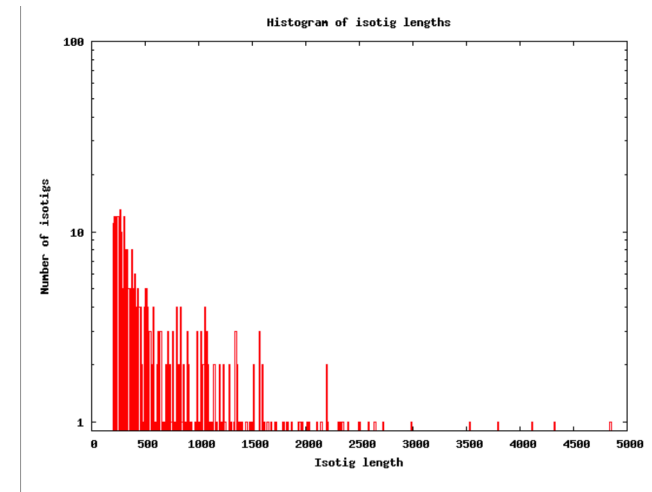
(--right)

Step 5: Assessing Quality of Assembly

Important statistics for assembly quality:

Contig Length Distribution

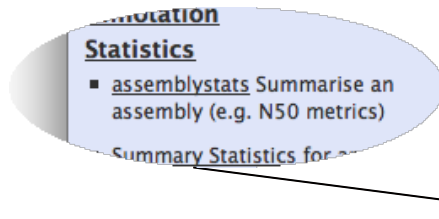
Assemblies will typically produce a number of complete contigs representing whole transcripts, and a large number of partial transcripts. This biases the average contig length toward the low end. The N50 is a measure weighted by total sequence length in the assembly.





Step 5: Assessing Quality of Assembly

Getting these stats in Galaxy:



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

The screenshot shows the Galaxy web interface at <https://galaxy.indiana.edu/galaxy-upgrade/root>. The main workspace displays the 'assemblystats (version 1.0.1)' tool. The 'Type of read:' dropdown is set to 'Isotig (if from transcriptomic assembly)'. The 'Source file in FASTA format:' dropdown is set to '84: Trinity on data 20 and data 21: Assembled Transcripts'. The 'Execute' button is visible. The right sidebar shows the history of runs, including '240: Sorted contigs' and '239: Assembly statistics'. The '239: Assembly statistics' run is expanded, showing a tabular output with the following statistics for isotig lengths:

Statistics for isotig lengths:
Min isotig length:
Max isotig length:
Mean isotig length:
Standard deviation of isotig length:
Median isotig length:



Step 5: Assessing Quality of Assembly

Type of read

Isotig (if from transcriptomic assembly)

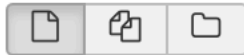
Is this from an genomic (contig) or transcriptomic assembly (isotig) or are these raw reads (read)

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



13: Trinity on data 10, data 6, and others: Assembled Transcripts

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



Step 8: Differential Expression

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

htseq-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

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

RSEM_abundance_estimation run RSEM to estimate transcript abundances (Galaxy Version 0.0.1) Options

Transcripts Fasta




   15: Trinity on data 12, data 8, and others: Assembled Transcripts

Fasta sequences against which reads are aligned. This may be the Assembled Transcripts file from Trinity.




Paired or Single-end data?

Paired

Left/Forward strand reads


   7: Trimmomatic on Sp_ds.left.fq (R1 paired)

Right/Reverse strand reads

   8: Trimmomatic on Sp_ds.right.fq (R2 paired)

Strand-specific Library Type

None

 Execute

Here's an example for the DS dataset.

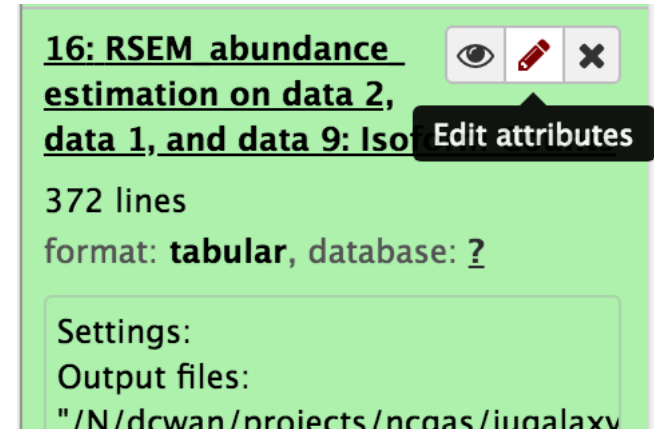
Rerun the job and change the two reads files to run the HS set as well.



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.



Edit attributes ↺ Auto-detect 💾 Save attributes

Name:



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

htseq-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 differential expression
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.






abundance_estimation_to_matrix Join RSEM estimates from multiple samples into a single matrix (Galaxy Version 0.0.1) Options

RSEM abundance estimates for samples

1: RSEM abundance estimates for samples

Add file






23: RSEM DS Gene counts

column label

ds

2: RSEM abundance estimates for samples

Add file



25: RSEM HS Gene counts

column label

hs

+ Insert RSEM abundance estimates for samples

Execute



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.

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

htseq-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

Use the Counts matrix and the Trinity assembled transcripts.

EdgeR_differentialExpression Identify Differentially Expressed Transcripts Using EdgeR Options
(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



Step 8: Differential Expression

The last step is to visualize the results from the statistical analysis.

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

htseq-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

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


Analyze_Differential_Expression Analyze differential expression (Galaxy Version 0.0.1) Options

EdgeR tar gz file
   20: EdgeR_differentialExpression on data 13 and data 19: EdgeR_Results.tar.gz

TMM Normalized FPKM matrix
   19: abundance_estimation_to_matrix on data 17 and data 15: TMM EXPR Matrix

P-value

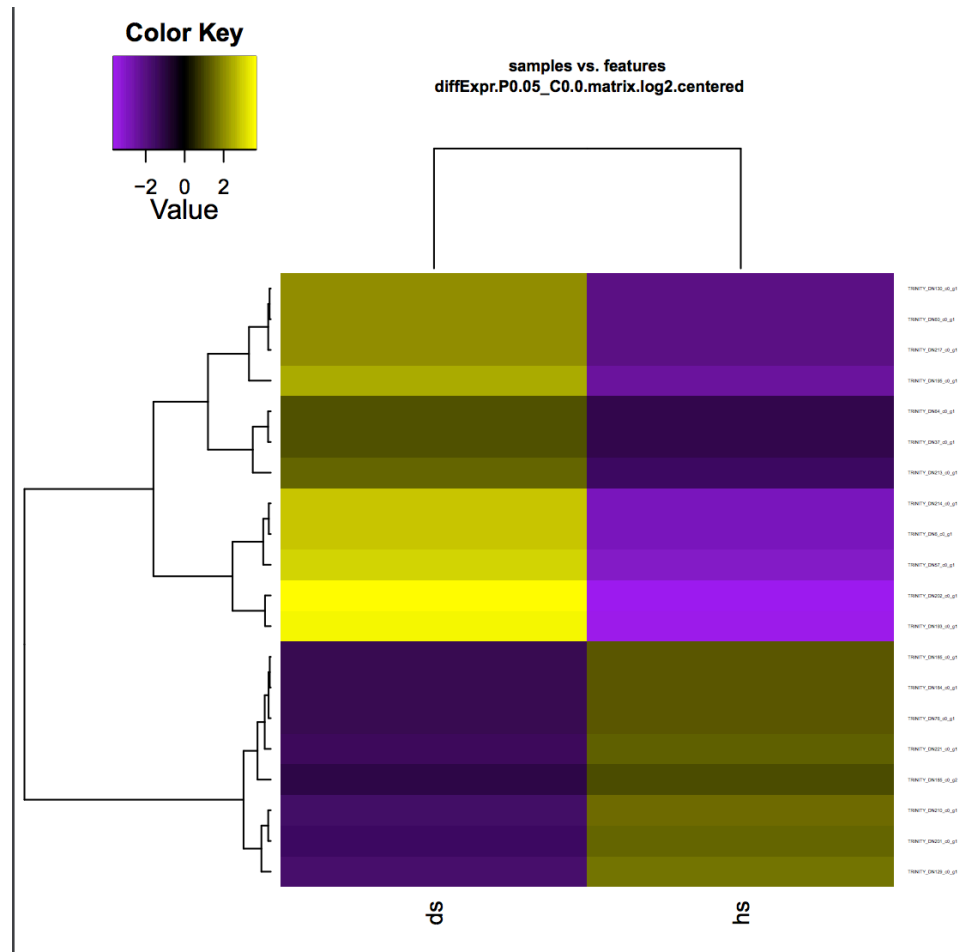
C-value

 **Execute**



Step 8: Differential Expression

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





INDIANA UNIVERSITY

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.



INDIANA UNIVERSITY

Fin

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