Intro to Using Galaxy
–
For Bioinformatics

Tom Doak

Carrie Ganote

National Center for Genome Analysis Support

September 17, 2013

INDIANA UNIVERSITY

Summary

- Who is NCGAS?
- Galaxy – what is it?
- Galaxy 101 – a guided tour
- Short intro to transcriptome assembly, as an example

Who is NCGAS?

The National Center for Genome Analysis Support is based at IU in Bloomington, but caters to a national audience with support from the NSF.

We provide computational resources and support for genomics, transcriptomics, and meta projects.

Our Services

NCGAS provides support in the form of long- and short-term consultation for genomics, proteomics, transcriptomics, and meta projects. We are happy to answer questions about software, methods, and pipelines; basic Linux use; experimental setup; and interpretation of results.

We administer bioinformatics software installation and upgrades on the Mason cluster at IU, as well as provide access to Mason to users of XSEDE's national infrastructure. We provide support letters for NSF proposals pledging our compute resources.

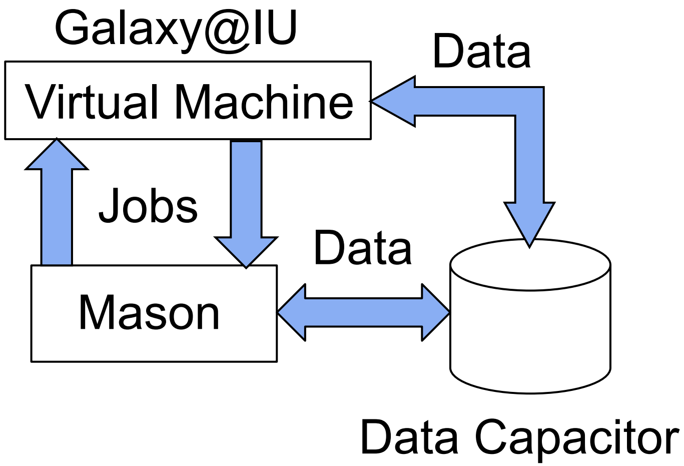
Last, but not least, we install and maintain the local Galaxy instances for Indiana University: IU, NCGAS, and Rockhopper.

What is Galaxy?

Galaxy is a web-based framework for running command-line utilities from a snazzy graphical user interface.

The Galaxy web server that we will be using today is hosted at Indiana University on the XSEDE virtual machines. This is a different “instance” than Galaxy Main, which is hosted at Penn State.

Our instance at IU



Galaxy@IU

Virtual Machine

Data

Jobs

Data

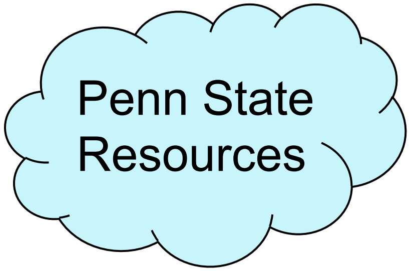
Mason

Data Capacitor

Why choose us:

- IU only – less busy!
- Large RAM jobs possible
- Custom tools on request
- On-site support

Galaxy Main

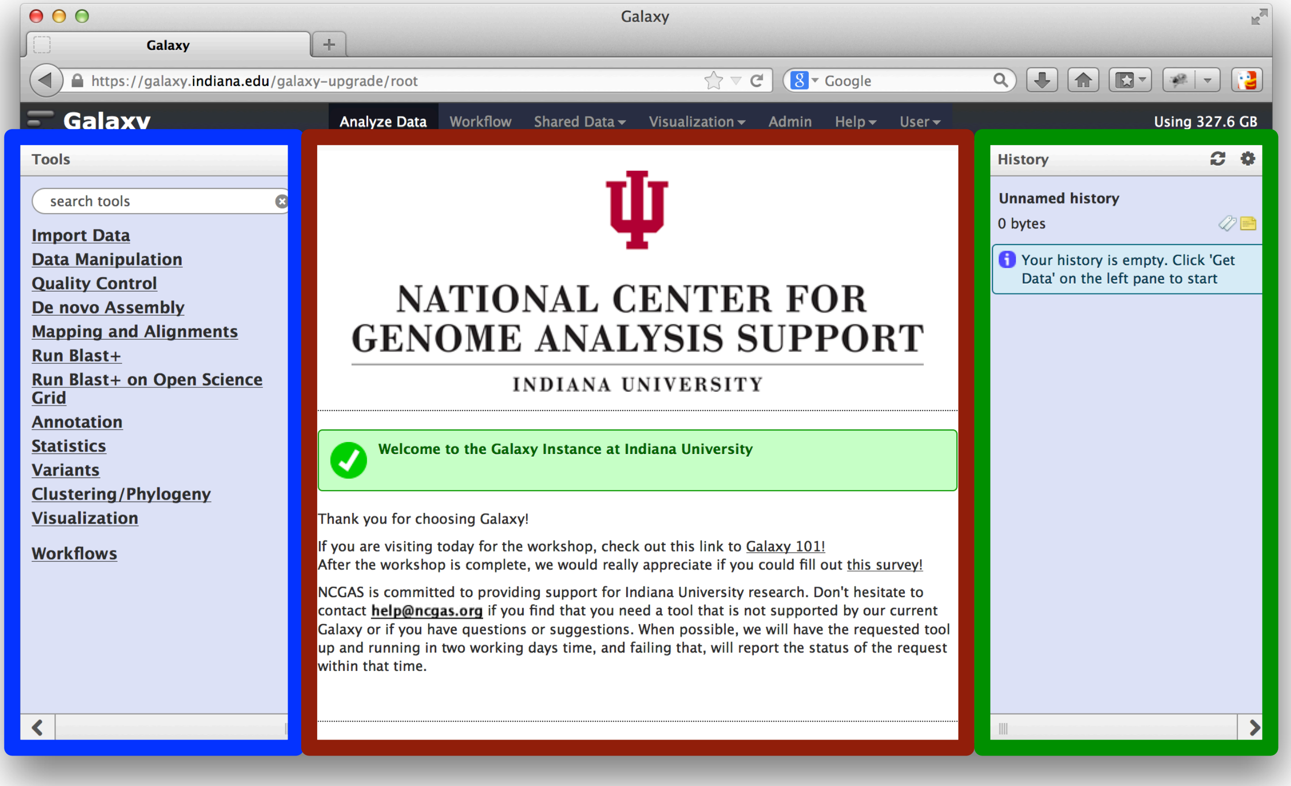


Penn State Resources

INDIANA UNIVERSITY

Galaxy Anatomy and Physiology

Tool bar – contains the available steps to apply to data



Your history is empty. Click 'Get Data' on the left pane to start

NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

Welcome to the Galaxy Instance at Indiana University

Thank you for choosing Galaxy!

If you are visiting today for the workshop, check out this link to Galaxy 101!
After the workshop is complete, we would really appreciate if you could fill out this survey!

NCGAS is committed to providing support for Indiana University research. Don't hesitate to contact help@ncgas.org if you find that you need a tool that is not supported by our current Galaxy or if you have questions or suggestions. When possible, we will have the requested tool up and running in two working days time, and failing that, will report the status of the request within that time.

History – shows steps previously taken to manipulate input data sets

Focus pane – shows options, parameters, and output for current item.

Galaxy 101 – Quick Start

We will depart this slideshow for a short time as we go through the basics of Galaxy using the Galaxy 101 tutorial. You can find a link to it on the home page for galaxy.indiana.edu.

You can choose to follow along either on IU Galaxy or on Galaxy Main – the tool layout is slightly different between the two instances.

Today's Menu Item

Cristobal Rojas, La miseria (1886) from Wikipedia.

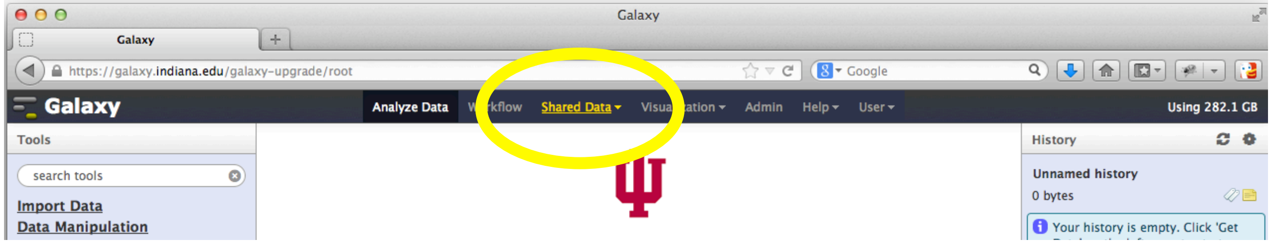
We will be assembling the DNA Polymerase protein units from the H37Rv strain of *Mycobacterium tuberculosis*, the causative agent of TB, also known as the consumption.

The raw reads originated from the Short Read Archive on NCBI. The accession number for the set is SRX212035.

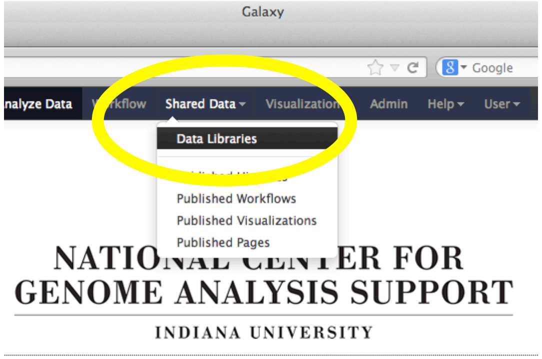
This dataset consists of paired-end, ~75bp RNA-Seq reads.

Let’s get some sequence data

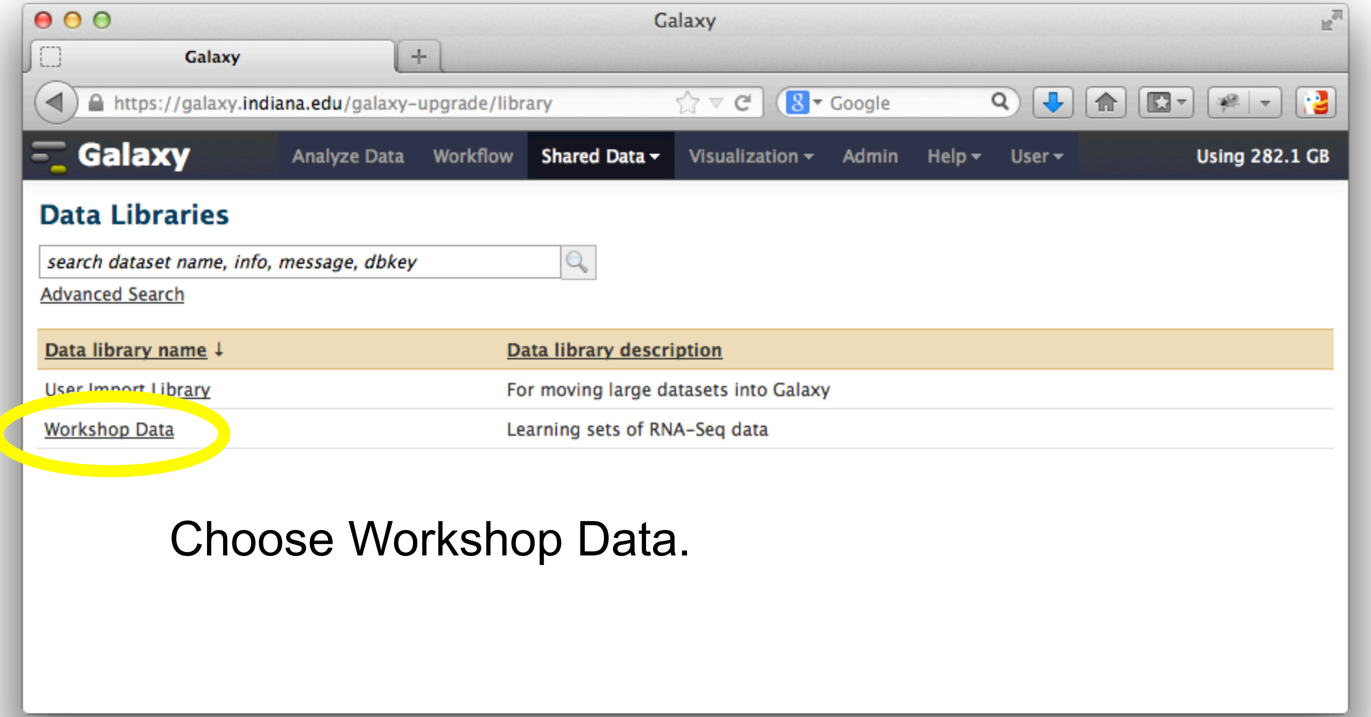
Galaxy allows users to publish their data to the entire user base.



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



Let’s get some sequence data



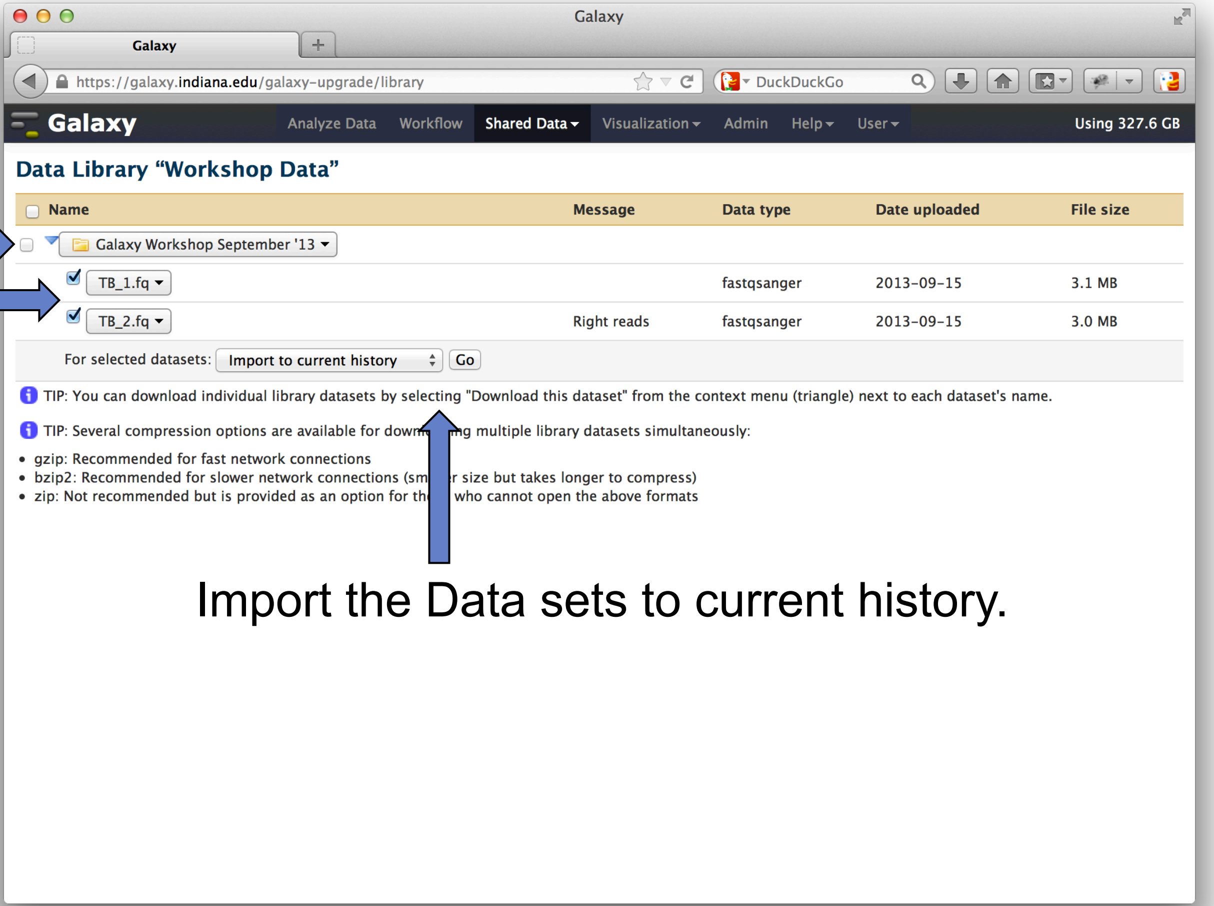
Data Libraries

search dataset name, info, message, dbkey

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.

Let’s get some sequence data



Galaxy

https://galaxy.indiana.edu/galaxy-upgrade/library

Data Library “Workshop Data”

□ Name	Message	Data type	Date uploaded	File size
□ Galaxy Workshop September '13 ▾				
☑ TB_1.fq ▾		fastqsanger	2013-09-15	3.1 MB
☑ TB_2.fq ▾	Right reads	fastqsanger	2013-09-15	3.0 MB

For selected datasets: Import to current history [Go]

TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle) next to each dataset's name.

TIP: Several compression options are available for downloading multiple library datasets simultaneously:

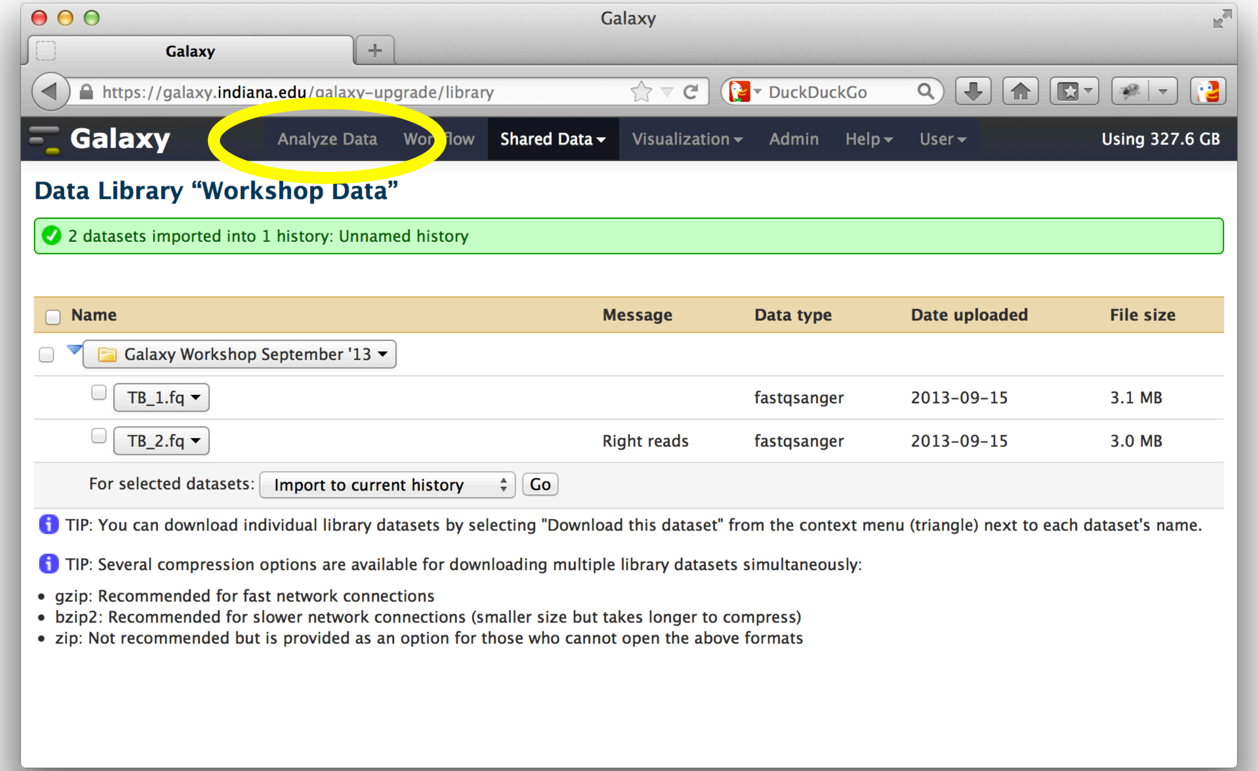
- gzip: Recommended for fast network connections
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)
- zip: Not recommended but is provided as an option for those who cannot open the above formats

Expand folder

Check both boxes

Import the Data sets to current history.

Let’s get some sequence data



Data Library “Workshop Data”

✓ 2 datasets imported into 1 history: Unnamed history

□ Name	Message	Data type	Date uploaded	File size
□ Galaxy Workshop September '13 ▾				
□ TB_1.fq ▾		fastqsanger	2013-09-15	3.1 MB
□ TB_2.fq ▾	Right reads	fastqsanger	2013-09-15	3.0 MB

For selected datasets: Import to current history ⇕ Go

TIP: You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle) next to each dataset's name.

TIP: Several compression options are available for downloading multiple library datasets simultaneously:

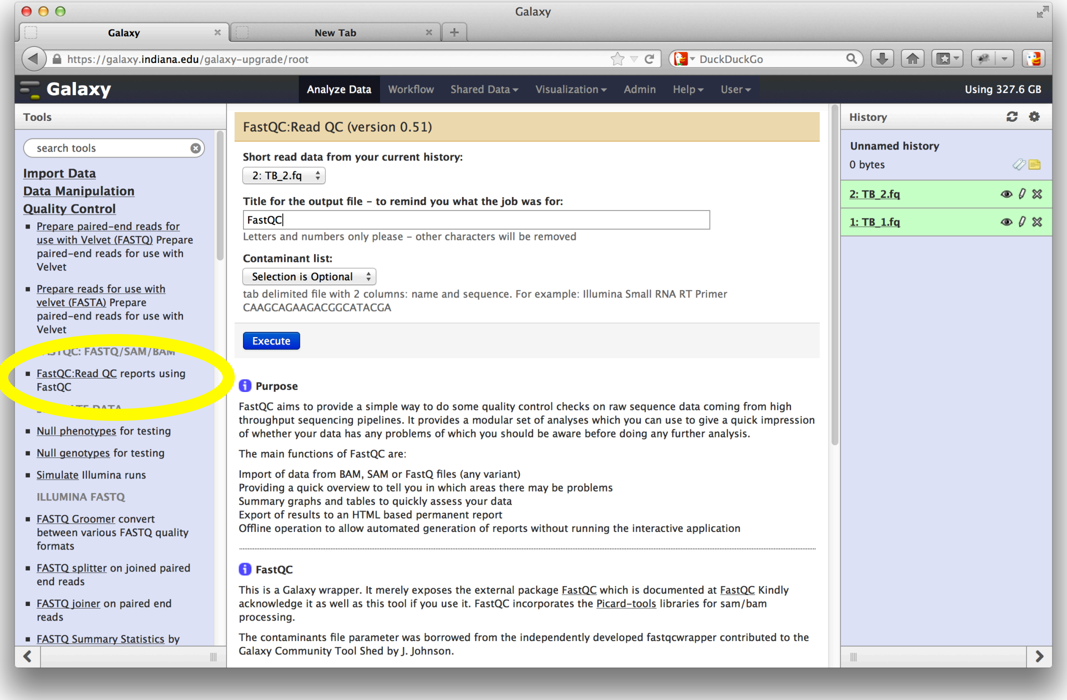
- gzip: Recommended for fast network connections
- bzip2: Recommended for slower network connections (smaller size but takes longer to compress)
- zip: Not recommended but is provided as an option for those who cannot open the above formats

Data set is imported – Click on Analyze Data to return.

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.



- Prepare paired-end reads for use with Velvet (FASTQ) Prepare paired-end reads for use with Velvet
- Prepare reads for use with velvet (FASTA) Prepare paired-end reads for use with Velvet
- FastQC:Read QC reports using FastQC
- Null phenotypes for testing
- Null genotypes for testing
- Simulate Illumina runs
- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by

FastQC:Read QC (version 0.51)

Short read data from your current history: 2: TB_2.fq

Title for the output file – to remind you what the job was for: FastQC

Letters and numbers only please – other characters will be removed

Contaminant list: Selection is Optional

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

Purpose

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

Import of data from BAM, SAM or FastQ files (any variant)
Providing a quick overview to tell you in which areas there may be problems
Summary graphs and tables to quickly assess your data
Export of results to an HTML based permanent report
Offline operation to allow automated generation of reports without running the interactive application

FastQC

This is a Galaxy wrapper. It merely exposes the external package FastQC which is documented at FastQC Kindly acknowledge it as well as this tool if you use it. FastQC incorporates the Picard-tools libraries for sam/bam processing.

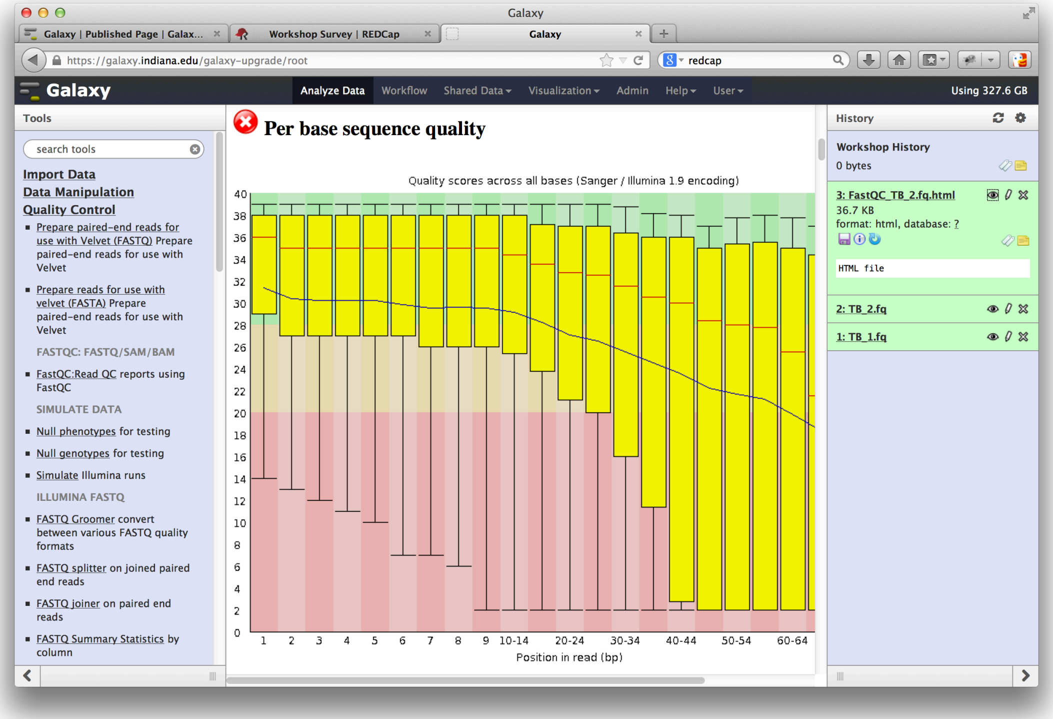
The contaminants file parameter was borrowed from the independently developed fastqcwrapper contributed to the Galaxy Community Tool Shed by J. Johnson.

Choose either the left or right reads. Compare the results with your neighbor.

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.



Import Data
Data Manipulation
Quality Control

- Prepare paired-end reads for use with Velvet (FASTQ) Prepare paired-end reads for use with Velvet
- Prepare reads for use with velvet (FASTA) Prepare paired-end reads for use with Velvet

FASTQC: FASTQ/SAM/BAM

- FastQC:Read QC reports using FastQC

SIMULATE DATA

- Null phenotypes for testing
- Null genotypes for testing
- Simulate Illumina runs

ILLUMINA FASTQ

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

Per base sequence quality

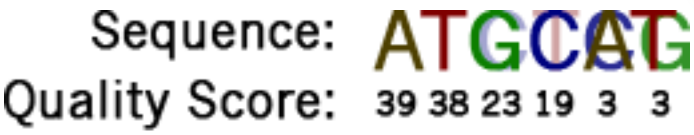
Workshop History
0 bytes

3: FastQC_TB_2.fq.html
36.7 KB
format: html, database: ?

```
HTML file
```

2: TB_2.fq

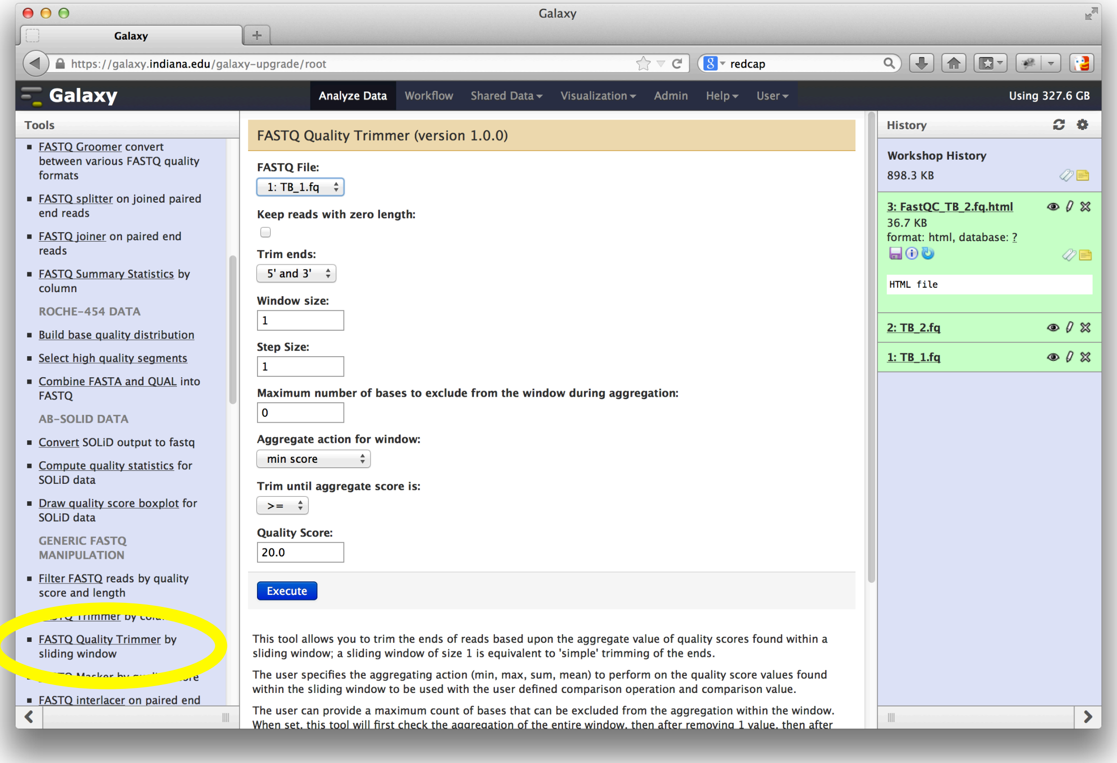
1: TB_1.fq



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 the FASTQ quality trimmer by sliding window to trim reads based on quality score.



- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score
- FASTQ interlacer on paired end

FASTQ Quality Trimmer (version 1.0.0)

FASTQ File: 1: TB_1.fq

Keep reads with zero length: □

Trim ends: 5' and 3'

Window size: 1

Step Size: 1

Maximum number of bases to exclude from the window during aggregation: 0

Aggregate action for window: min score

Trim until aggregate score is: >=

Quality Score: 20.0

Execute

This tool allows you to trim the ends of reads based upon the aggregate value of quality scores found within a sliding window; a sliding window of size 1 is equivalent to 'simple' trimming of the ends.

The user specifies the aggregating action (min, max, sum, mean) to perform on the quality score values found within the sliding window to be used with the user defined comparison operation and comparison value.

The user can provide a maximum count of bases that can be excluded from the aggregation within the window. When set, this tool will first check the aggregation of the entire window, then after removing 1 value, then after

Workshop History
898.3 KB

3: FastQC_TB_2.fq.html
36.7 KB
format: html, database: ?
HTML file

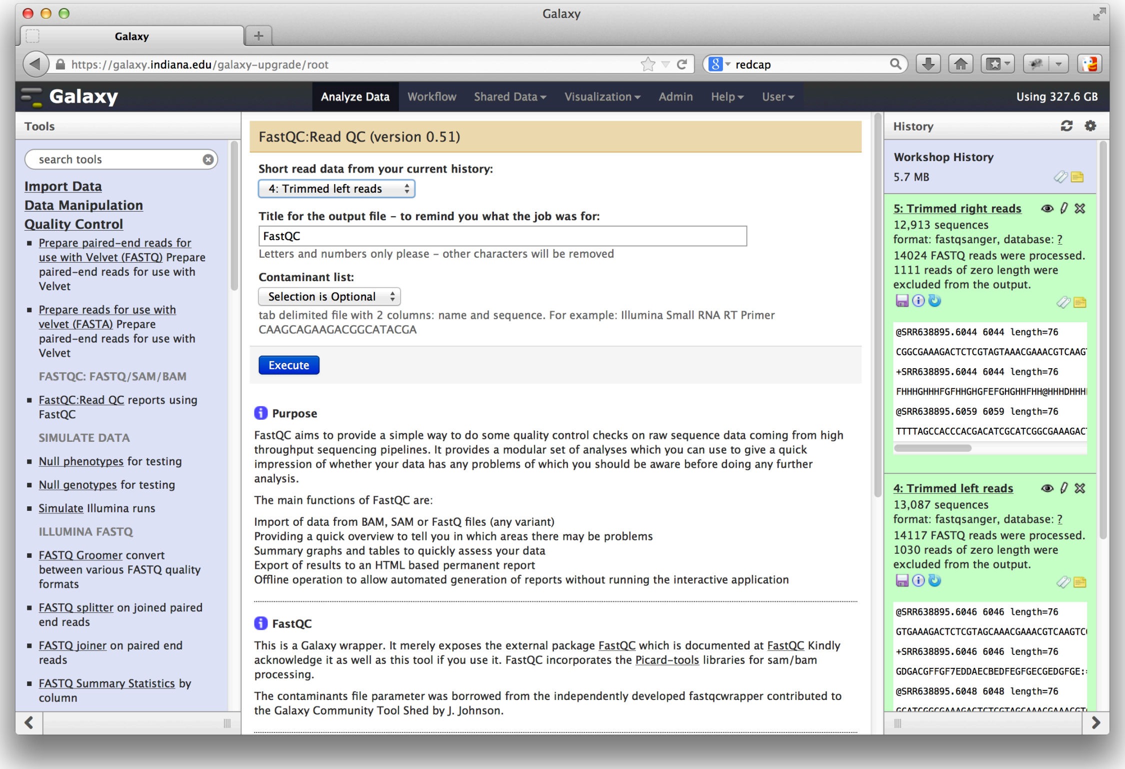
2: TB_2.fq

1: TB_1.fq

Run this tool for both input data sets.

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.



- Prepare paired-end reads for use with Velvet (FASTQ) Prepare paired-end reads for use with Velvet
- Prepare reads for use with velvet (FASTA) Prepare paired-end reads for use with Velvet

FASTQC: FASTQ/SAM/BAM

- FastQC:Read QC reports using FastQC

SIMULATE DATA

- Null phenotypes for testing
- Null genotypes for testing
- Simulate Illumina runs

ILLUMINA FASTQ

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

FastQC:Read QC (version 0.51)

Short read data from your current history: 4: Trimmed left reads

Title for the output file - to remind you what the job was for: FastQC
Letters and numbers only please - other characters will be removed

Contaminant list: Selection is Optional
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

Purpose

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

The main functions of FastQC are:

Import of data from BAM, SAM or FastQ files (any variant)
Providing a quick overview to tell you in which areas there may be problems
Summary graphs and tables to quickly assess your data
Export of results to an HTML based permanent report
Offline operation to allow automated generation of reports without running the interactive application

FastQC

This is a Galaxy wrapper. It merely exposes the external package FastQC which is documented at FastQC Kindly acknowledge it as well as this tool if you use it. FastQC incorporates the Picard-tools libraries for sam/bam processing.

The contaminants file parameter was borrowed from the independently developed fastqcwrapper contributed to the Galaxy Community Tool Shed by J. Johnson.

Workshop History
5.7 MB

5: Trimmed right reads
12,913 sequences
format: fastqsanger, database: ?
14024 FASTQ reads were processed. 1111 reads of zero length were excluded from the output.

```
@SRR638895.6044 6044 length=76
CGGCGAAAGACTCTCGTAGTAAACGAAACGTCAAG
+SRR638895.6044 6044 length=76
FHHHGHHHFGFHHGHGFEFGHGHHFHH@HHHDHHH
@SRR638895.6059 6059 length=76
TTTTAGCCACCCACGACATCGCATCGGCGAAAGAC
```

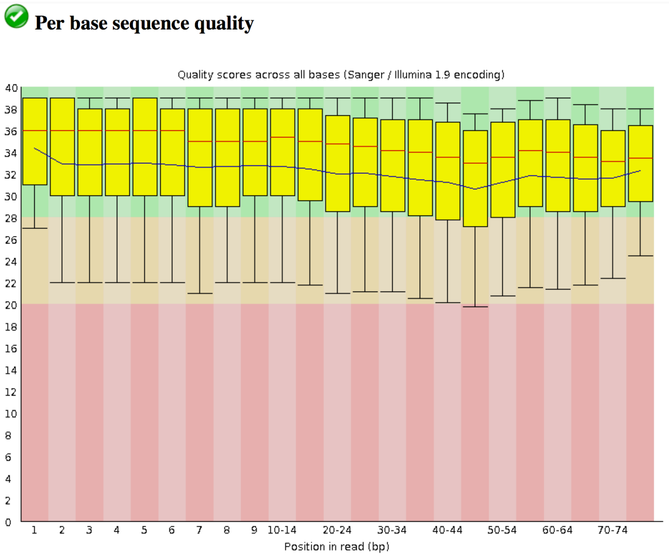
4: Trimmed left reads
13,087 sequences
format: fastqsanger, database: ?
14117 FASTQ reads were processed. 1030 reads of zero length were excluded from the output.

```
@SRR638895.6046 6046 length=76
GTGAAAGACTCTCGTAGCAAACGAAACGTCAAGTC
+SRR638895.6046 6046 length=76
GDGACGFFGF7EDDAECBEDFEGFGECGEDGFGE:
@SRR638895.6048 6048 length=76
GCATCGGCGAAAGACTCTCGTAGCAAACGAAACGT
```

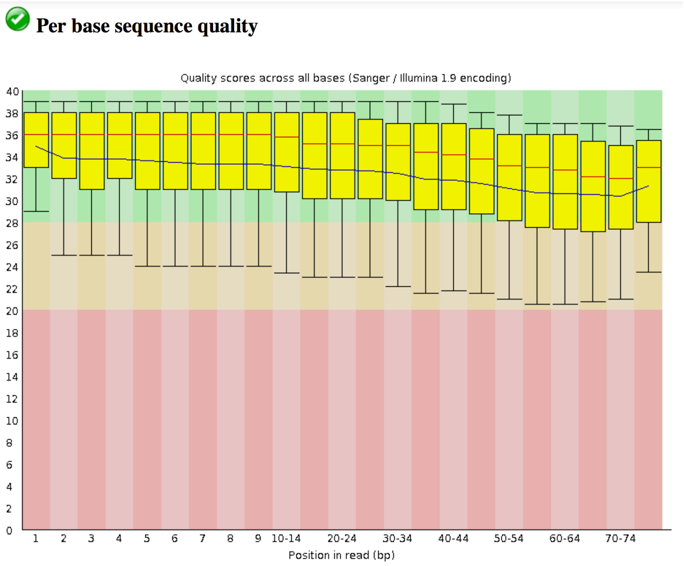
I renamed my trimmed files to help me keep them straight.

Step 3: Rinse, Repeat

Pictured are the left and right reads after trimming is complete. These will do!



✓ **Per base sequence quality**

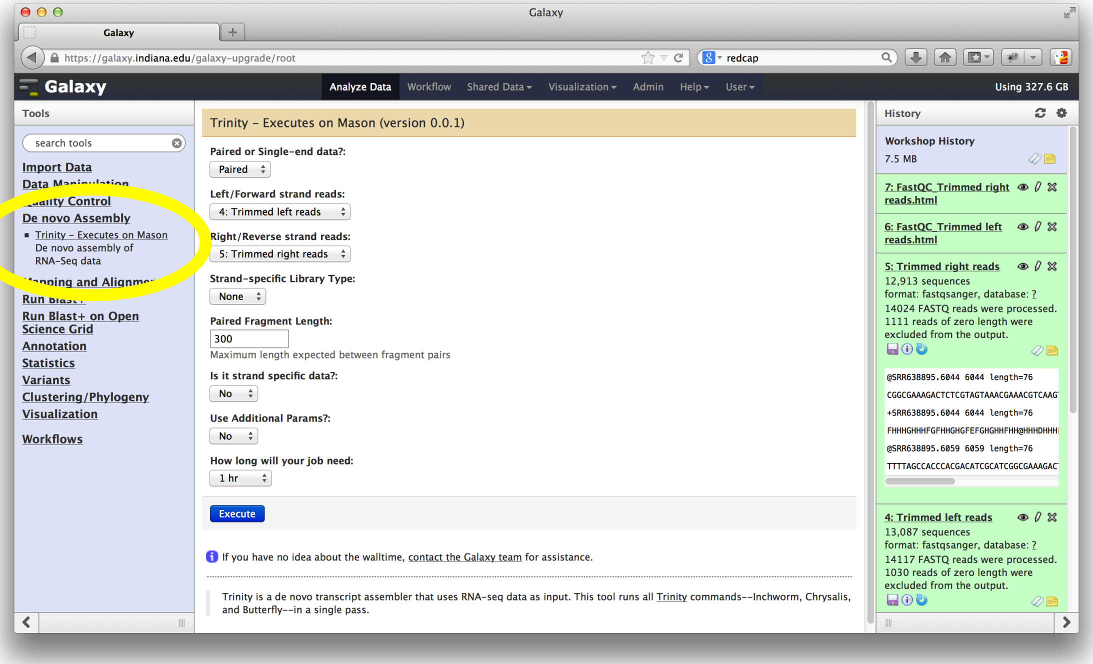


✓ **Per base sequence quality**

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 has been optimized for use on Mason. We will use it to assemble our reads.



Trinity – Executes on Mason (version 0.0.1)

Paired or Single-end data?: Paired

Left/Forward strand reads: 4: Trimmed left reads

Right/Reverse strand reads: 5: Trimmed right reads

Strand-specific Library Type: None

Paired Fragment Length: 300

Maximum length expected between fragment pairs

Is it strand specific data?: No

Use Additional Params?: No

How long will your job need: 1 hr

Execute

If you have no idea about the walltime, contact the Galaxy team for assistance.

Trinity is a de novo transcript assembler that uses RNA-seq data as input. This tool runs all Trinity commands--Inchworm, Chrysalis, and Butterfly--in a single pass.

Workshop History
7.5 MB

7: FastQC_Trimmed right reads.html

6: FastQC_Trimmed left reads.html

5: Trimmed right reads
12,913 sequences
format: fastqsanger, database: ?
14024 FASTQ reads were processed.
1111 reads of zero length were excluded from the output.

```
@SRR638895.6044 6044 length=76
CGGCGAAAGACTCTCGTAGTAAACGAAACGTCAAG
+SRR638895.6044 6044 length=76
FHHHGHHHFGFHHGHGFEFGHGHHFHH@HHHDHHH
@SRR638895.6059 6059 length=76
TTTTAGCCACCCACGACATCGCATCGGCGAAAGAC
```

4: Trimmed left reads
13,087 sequences
format: fastqsanger, database: ?
14117 FASTQ reads were processed.
1030 reads of zero length were excluded from the output.

It finished! We're done, right?

An assembler solves a computer problem of putting together a puzzle from tiny pieces. The output of the assembler is a guess – but we don't know how accurate it is. We could look at:

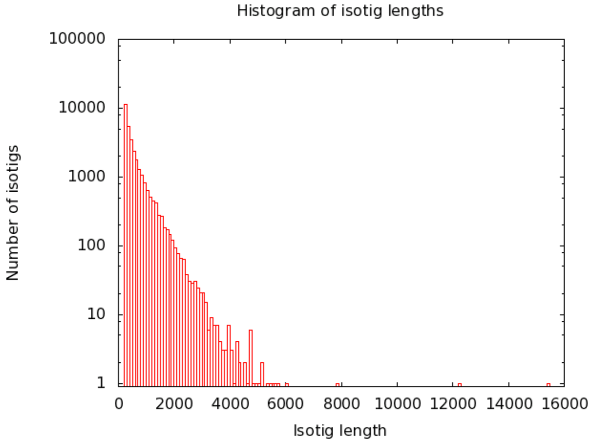
- Basic stats of the assembly – "Contigs"
  - Number of "Contigs" vs. Expected Number
  - N50 – a weighted average
  - Average Length
  - Max Length
- Check contigs against known genes with Blast (large or rare transcripts)

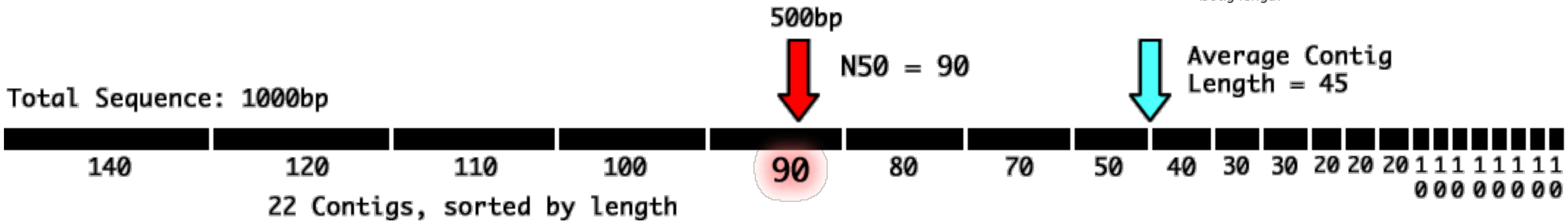
Step 5: Assessing Quality of Assembly

Important statistics for assembly quality:

Contig Length Distribution

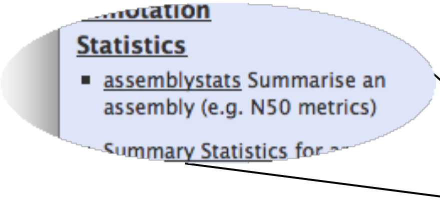
Assemblies will typically produce a number of complete contigs representing whole transcripts, and a large number of partial transcripts. This biases the average contig length toward the low end. The N50 is a measure weighted by total sequence length in the assembly.



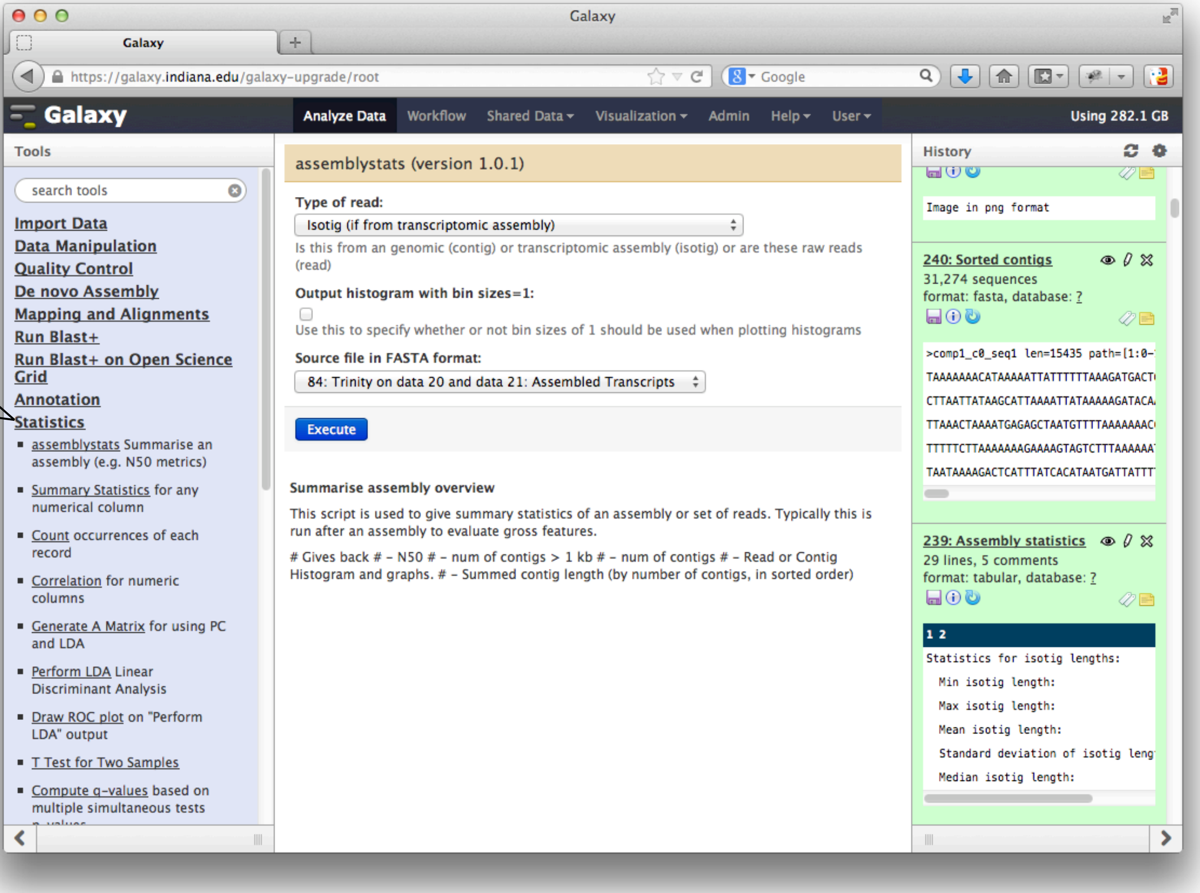


Step 5: Assessing Quality of Assembly

Getting these stats in Galaxy:


...notation
Statistics
▪ assemblystats Summarise an assembly (e.g. N50 metrics)
Summary Statistics for ...

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



- assemblystats Summarise an assembly (e.g. N50 metrics)
- Summary Statistics for any numerical column
- Count occurrences of each record
- Correlation for numeric columns
- Generate A Matrix for using PC and LDA
- Perform LDA Linear Discriminant Analysis
- Draw ROC plot on "Perform LDA" output
- T Test for Two Samples
- Compute q-values based on multiple simultaneous tests p-values

assemblystats (version 1.0.1)

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: □

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

Source file in FASTA format: 84: Trinity on data 20 and data 21: Assembled Transcripts

Execute

Summarise assembly overview

This script is used to give summary statistics of an assembly or set of reads. Typically this is run after an assembly to evaluate gross features.

Gives back # – N50 # – num of contigs > 1 kb # – num of contigs # – Read or Contig Histogram and graphs. # – Summed contig length (by number of contigs, in sorted order)

```
Image in png format
```

240: Sorted contigs
31,274 sequences
format: fasta, database: ?

```
>comp1_c0_seq1 len=15435 path=[1:0-
TAAAAAAACATAAAAATTATTTTTTAAAGATGACT
CTTAATTATAAGCATTAAAATTATAAAAAGATACA
TTAAACTAAAATGAGAGCTAATGTTTTAAAAAAAC
TTTTTCTTAAAAAAAGAAAAGTAGTCTTTAAAAAA
TAATAAAAGACTCATTTATCACATAATGATTATTT
```

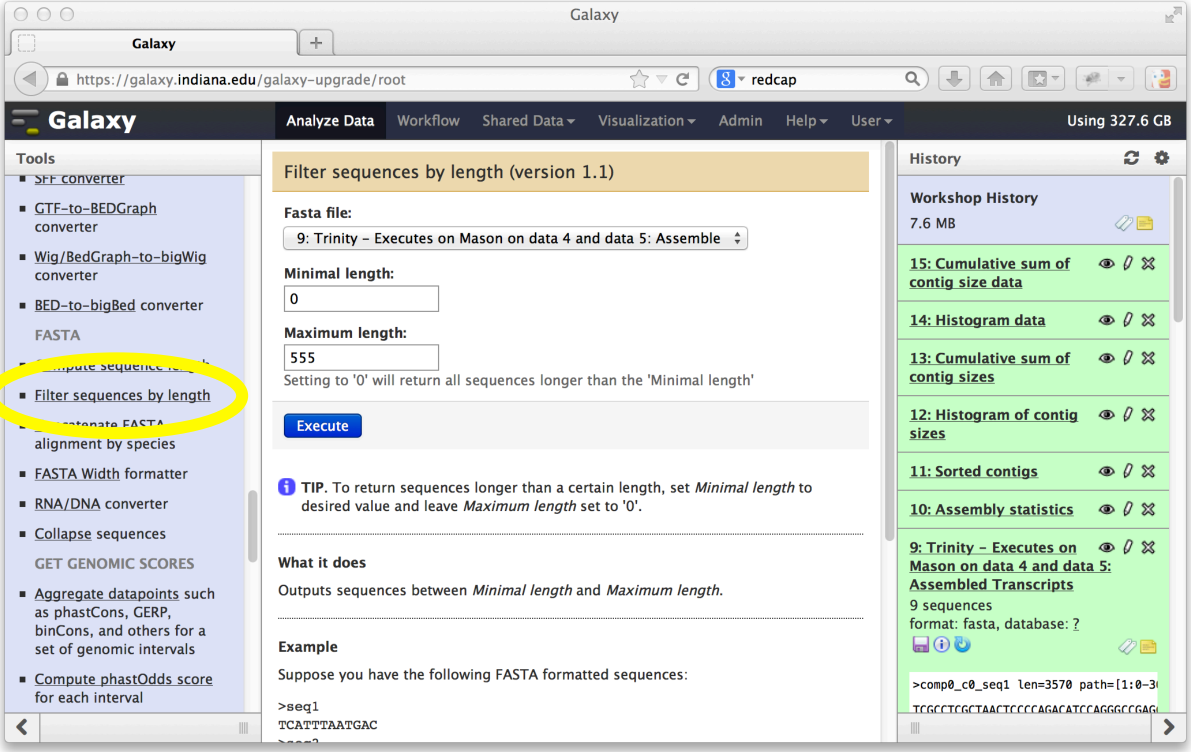
239: Assembly statistics
29 lines, 5 comments
format: tabular, database: ?

```
1 2
Statistics for isotig lengths:
  Min isotig length:
  Max isotig length:
  Mean isotig length:
  Standard deviation of isotig leng
  Median isotig length:
```

Step 6: Check Against Database

For this last step, we'll check to see how well our assembled transcripts compare to what we already know.

Use this step to give a rough annotation of genes, to make sure that your transcripts are from nuclear genes, or to gauge how complete your sequence is.


Galaxy
https://galaxy.indiana.edu/galaxy-upgrade/root
redcap
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 327.6 GB
Tools
SFF converter
GTF-to-BEDGraph converter
Wig/BedGraph-to-bigWig converter
BED-to-bigBed converter
FASTA
Compute sequence length
Filter sequences by length
Concatenate FASTA alignment by species
FASTA Width formatter
RNA/DNA converter
Collapse sequences
GET GENOMIC SCORES
Aggregate datapoints such as phastCons, GERP, binCons, and others for a set of genomic intervals
Compute phastOdds score for each interval
History
Workshop History
7.6 MB
15: Cumulative sum of contig size data
14: Histogram data
13: Cumulative sum of contig sizes
12: Histogram of contig sizes
11: Sorted contigs
10: Assembly statistics
9: Trinity – Executes on Mason on data 4 and data 5: Assembled Transcripts
9 sequences
format: fasta, database: ?
Execute

Filter sequences by length (version 1.1)

Fasta file: 9: Trinity – Executes on Mason on data 4 and data 5: Assemble

Minimal length: 0

Maximum length: 555

Setting to '0' will return all sequences longer than the 'Minimal length'

TIP. To return sequences longer than a certain length, set *Minimal length* to desired value and leave *Maximum length* set to '0'.

What it does

Outputs sequences between *Minimal length* and *Maximum length*.

Example

Suppose you have the following FASTA formatted sequences:

```
>seq1
TCATTTAATGAC
```

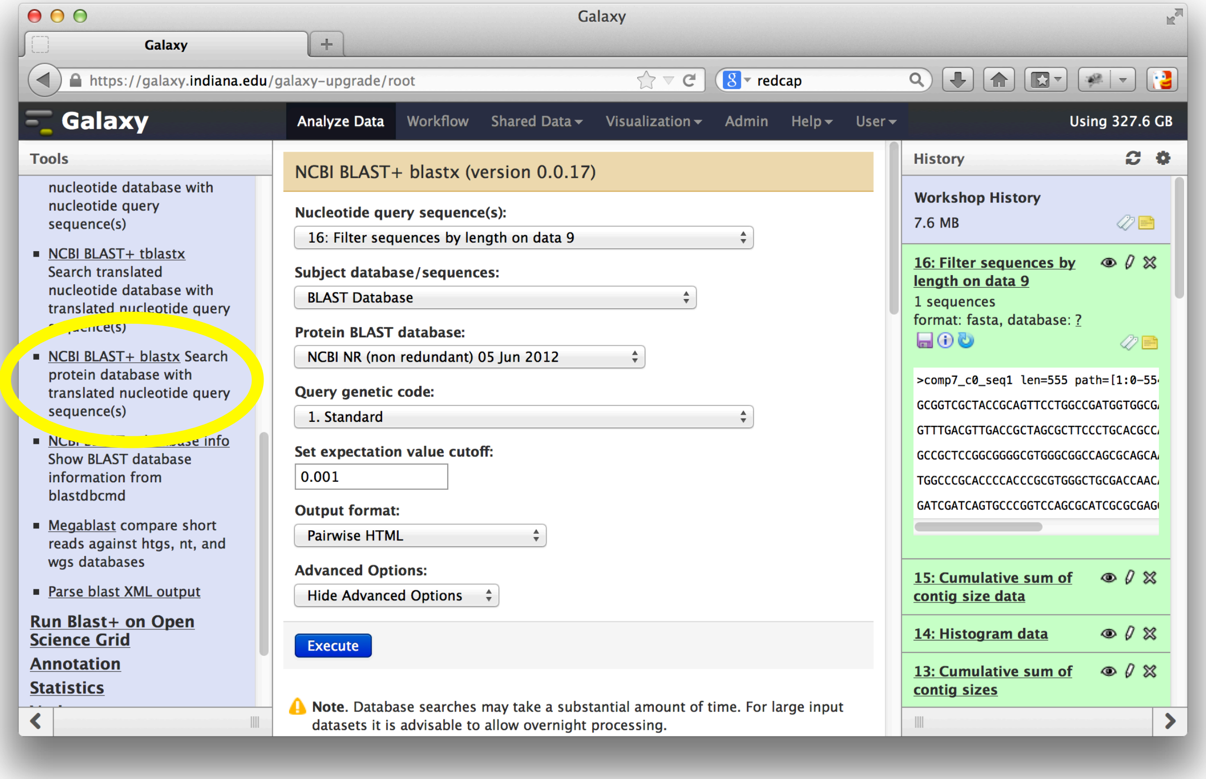
```
>comp0_c0_seq1 len=3570 path=[1:0-3
TCGCCTCGCTAACTCCCCAGACATCCAGGGCCGAG
```

For sake of time, we'll just Blast one gene. Filter out to get the smallest.

Step 6: Check Against Database

We will use Blastx to search the NR database for our gene.

Use default search settings for this test set.



nucleotide database with nucleotide query sequence(s)

- NCBI BLAST+ tblastx Search translated nucleotide database with translated nucleotide query sequence(s)
- NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)
- NCBI BLAST+ database info Show BLAST database information from blastdbcmd
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output

Run Blast+ on Open Science Grid

Annotation

Statistics

NCBI BLAST+ blastx (version 0.0.17)

Nucleotide query sequence(s): 16: Filter sequences by length on data 9

Subject database/sequences: BLAST Database

Protein BLAST database: NCBI NR (non redundant) 05 Jun 2012

Query genetic code: 1. Standard

Set expectation value cutoff: 0.001

Output format: Pairwise HTML

Advanced Options: Hide Advanced Options

Note. Database searches may take a substantial amount of time. For large input datasets it is advisable to allow overnight processing.

Workshop History
7.6 MB

16: Filter sequences by length on data 9
1 sequences
format: fasta, database: ?

```
>comp7_c0_seq1 len=555 path=[1:0-55
GCGGTCGCTACCGCAGTTCCTGGCCGATGGTGGCG
GTTTGACGTTGACCGCTAGCGCTTCCCTGCACGCC
GCCGCTCCGGCGGGGCGTGGGCGGCCAGCGCAGCA
TGGCCCGCACCCCACCCGCGTGGGCTGCGACCAAC
GATCGATCAGTGCCCGGTCCAGCGCATCGCGCGAG
```

15: Cumulative sum of contig size data

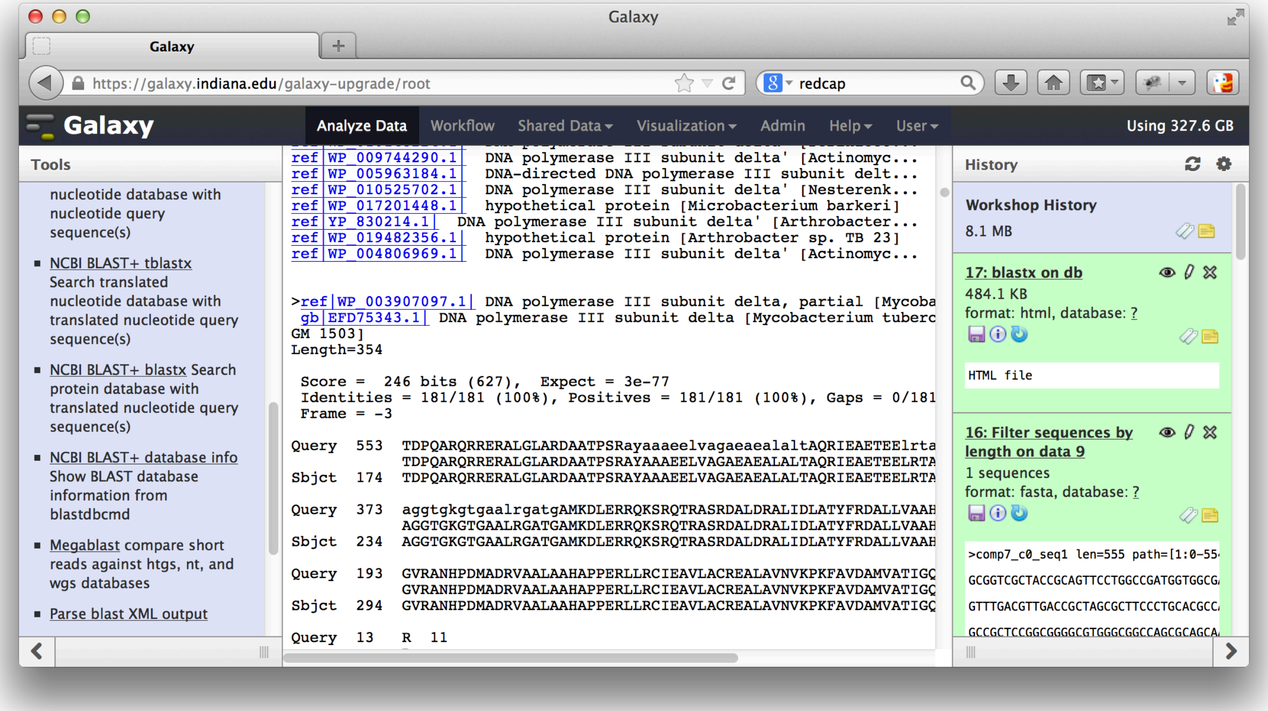
14: Histogram data

13: Cumulative sum of contig sizes

Make sure to choose Pairwise HTML output for readability.

Step 6: Check Against Database

We see the expected genes as the top hits!



nucleotide database with nucleotide query sequence(s)

- NCBI BLAST+ tblastx Search translated nucleotide database with translated nucleotide query sequence(s)
- NCBI BLAST+ blastx Search protein database with translated nucleotide query sequence(s)
- NCBI BLAST+ database info Show BLAST database information from blastdbcmd
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output

```
ref|WP_009744290.1|  DNA polymerase III subunit delta' [Actinomyc...
ref|WP_005963184.1|  DNA-directed DNA polymerase III subunit delt...
ref|WP_010525702.1|  DNA polymerase III subunit delta' [Nesterenk...
ref|WP_017201448.1|  hypothetical protein [Microbacterium barkeri]
ref|YP_830214.1|  DNA polymerase III subunit delta' [Arthrobacter...
ref|WP_019482356.1|  hypothetical protein [Arthrobacter sp. TB 23]
ref|WP_004806969.1|  DNA polymerase III subunit delta' [Actinomyc...


>ref|WP_003907097.1| DNA polymerase III subunit delta, partial [Mycoba
 gb|EFD75343.1| DNA polymerase III subunit delta [Mycobacterium tuberc
GM 1503]
Length=354

 Score =  246 bits (627),  Expect = 3e-77
 Identities = 181/181 (100%), Positives = 181/181 (100%), Gaps = 0/181
 Frame = -3

Query  553  TDPQARQRRERALGLARDAATPSRayaaaeelvagaeaealaltAQRIEAETEElrta
            TDPQARQRRERALGLARDAATPSRAYAAAEELVAGAEAEALALTAQRIEAETEELRTA
Sbjct  174  TDPQARQRRERALGLARDAATPSRAYAAAEELVAGAEAEALALTAQRIEAETEELRTA

Query  373  aggtgkgtgaalrgatgAMKDLERRQKSRQTRASRDALDRALIDLATYFRDALLVAAH
            AGGTGKGTGAALRGATGAMKDLERRQKSRQTRASRDALDRALIDLATYFRDALLVAAH
Sbjct  234  AGGTGKGTGAALRGATGAMKDLERRQKSRQTRASRDALDRALIDLATYFRDALLVAAH

Query  193  GVRANHPDMADRVAALAAHAPPERLLRCIEAVLACREALAVNVKPKFAVDAMVATIGQ
            GVRANHPDMADRVAALAAHAPPERLLRCIEAVLACREALAVNVKPKFAVDAMVATIGQ
Sbjct  294  GVRANHPDMADRVAALAAHAPPERLLRCIEAVLACREALAVNVKPKFAVDAMVATIGQ

Query  13   R  11
```

Workshop History
8.1 MB

17: blastx on db
484.1 KB
format: html, database: ?

```
HTML file
```

16: Filter sequences by length on data 9
1 sequences
format: fasta, database: ?

```
>comp7_c0_seq1 len=555 path=[1:0-55
GCGGTCGCTACCGCAGTTCCTGGCCGATGGTGGCG
GTTTGACGTTGACCGCTAGCGCTTCCCTGCACGCC
GCCGCTCCGGCGGGGCGTGGGCGGCCAGCGCAGCA
```

We could limit the number of hits depending on output desired.

Step 7..?

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
- 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