Inclass Activities Documentation

Release 170218

Asela Wijeratne

Nov 05, 2019

Bioinformatics part I

CHAPTER 1

Multiple Sequence Alignments

1.1 Getting started with MEGA

Note: There is an excellent tutorial on the MEGA site and this is excerpt of the tutorial for the exercise.

1.1.1 How to make an alignment using MEGA

Step 1 Open MEGA software and you will see a screen like in the following figure:

Step 2 Click on the small arrow on the "Align" tab

Step 3 Click on 'Edit/Build alignment'

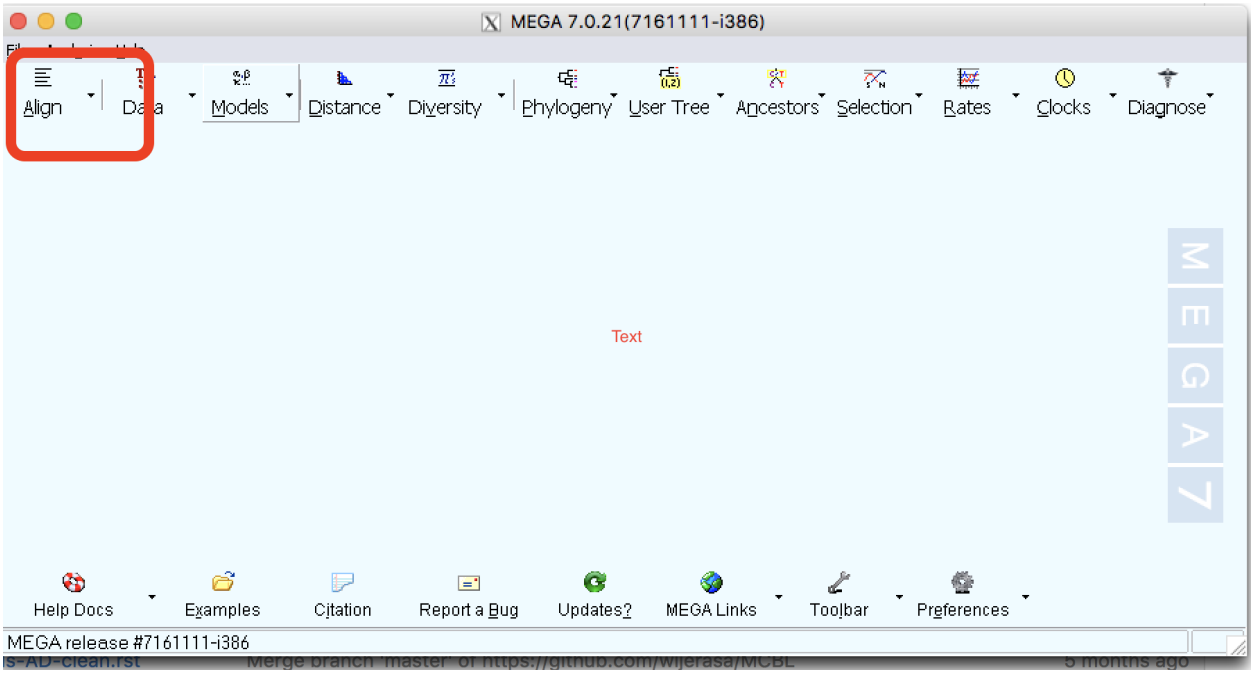
Step 4 Select a new alignment.

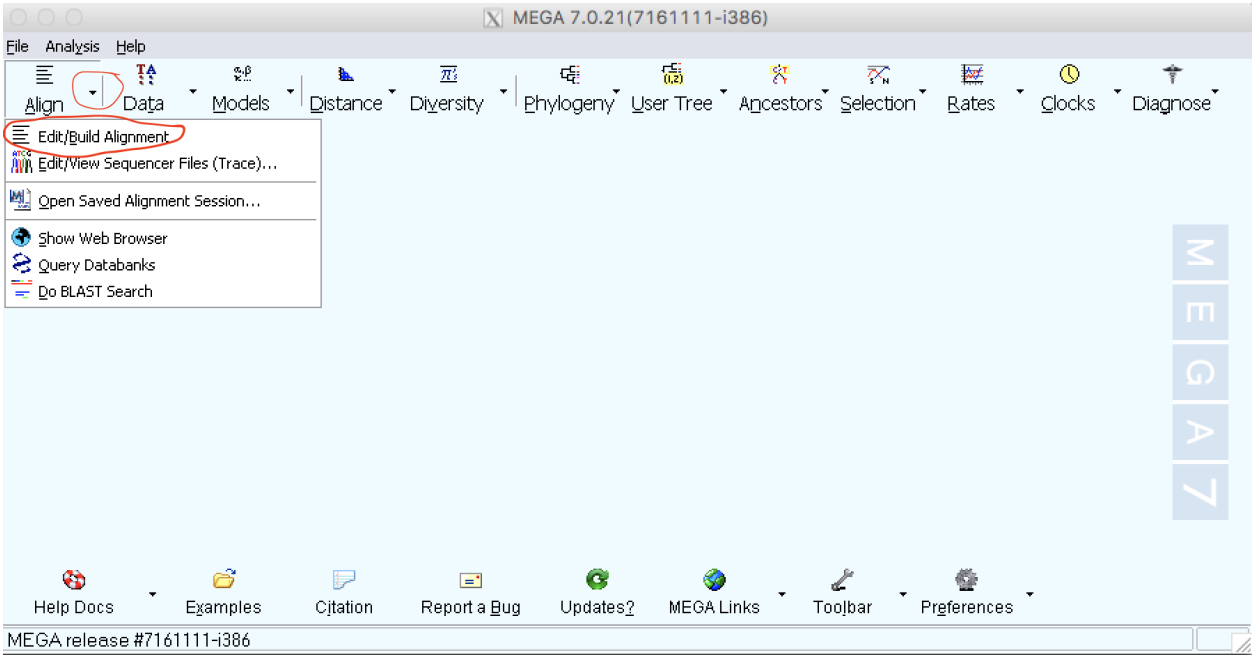
Step 4 Select protein

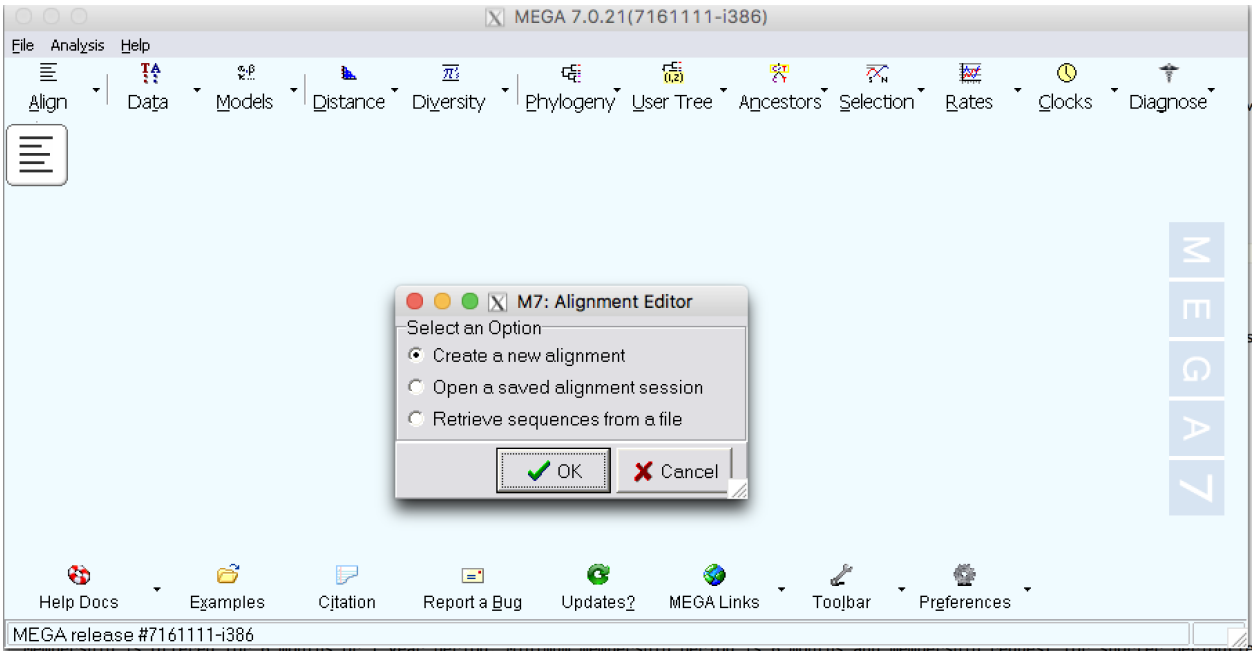
Step 5 From file open, select "seq_align2.fasta" and open file.

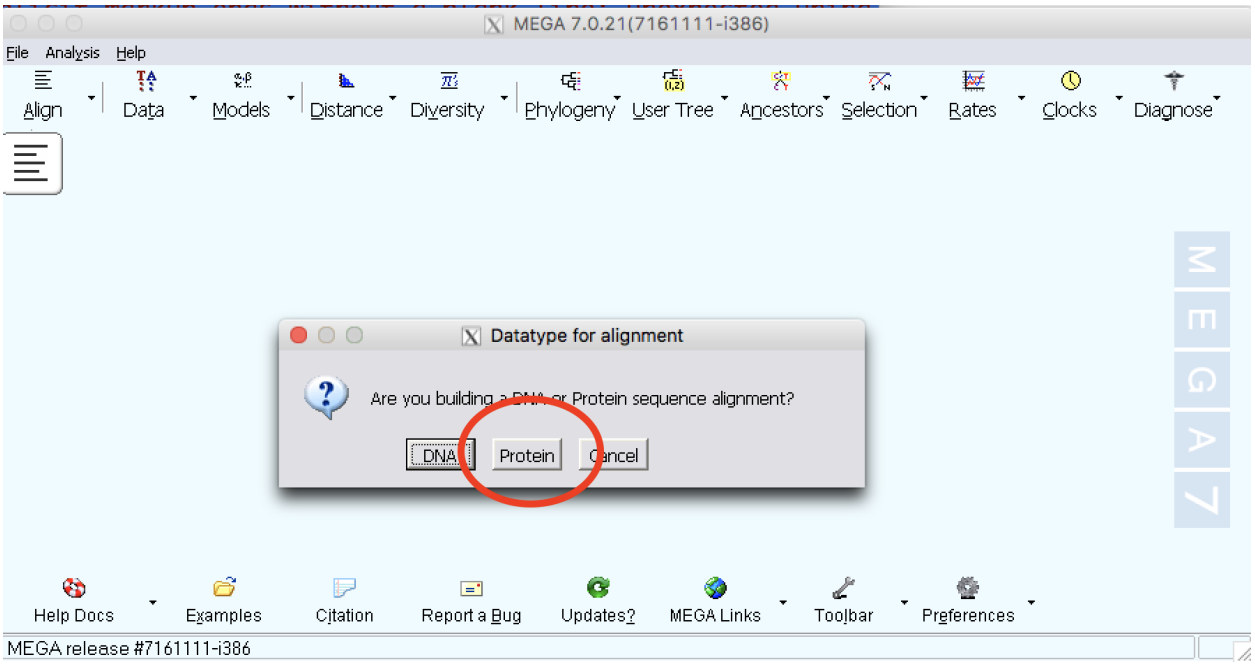
Step 6 From Edit, select all sequences. To do an alignment using Muscle, click on Muscle tab.

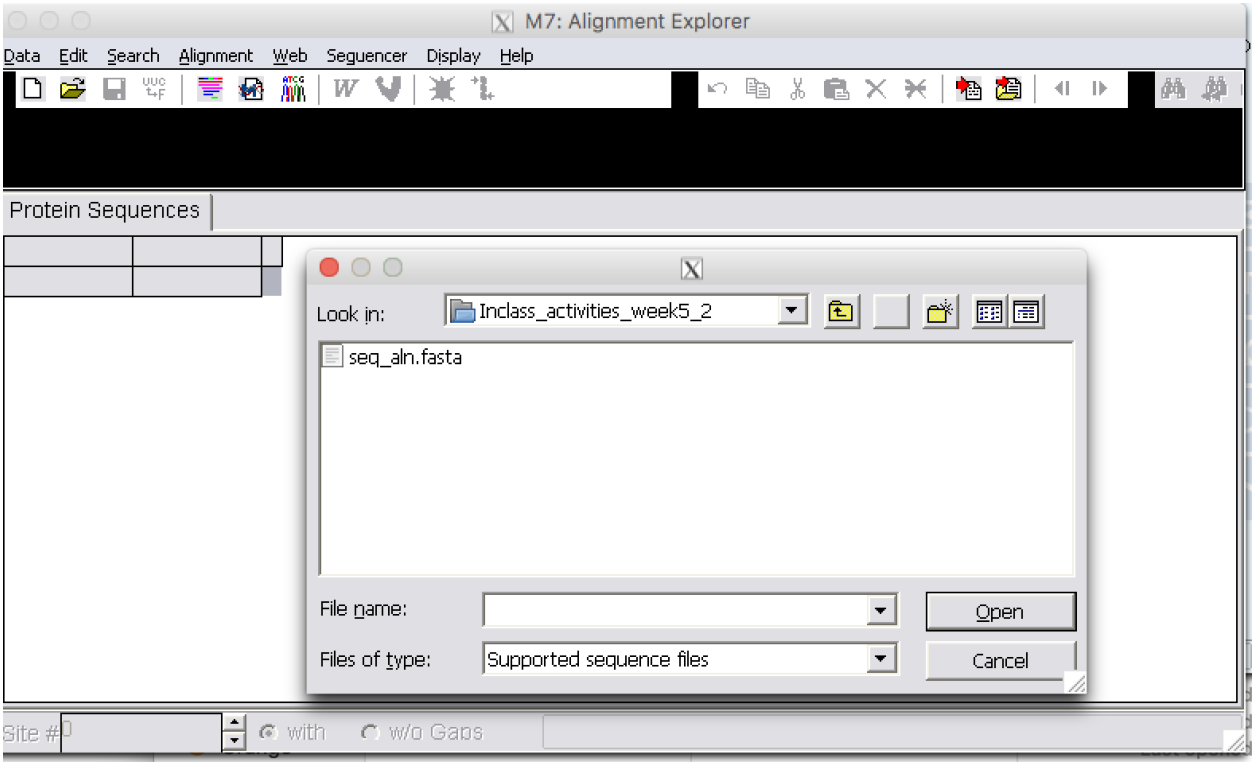
Step 7 Use default options and perform an alignment. To learn more about the options, go to the MEGA manual.

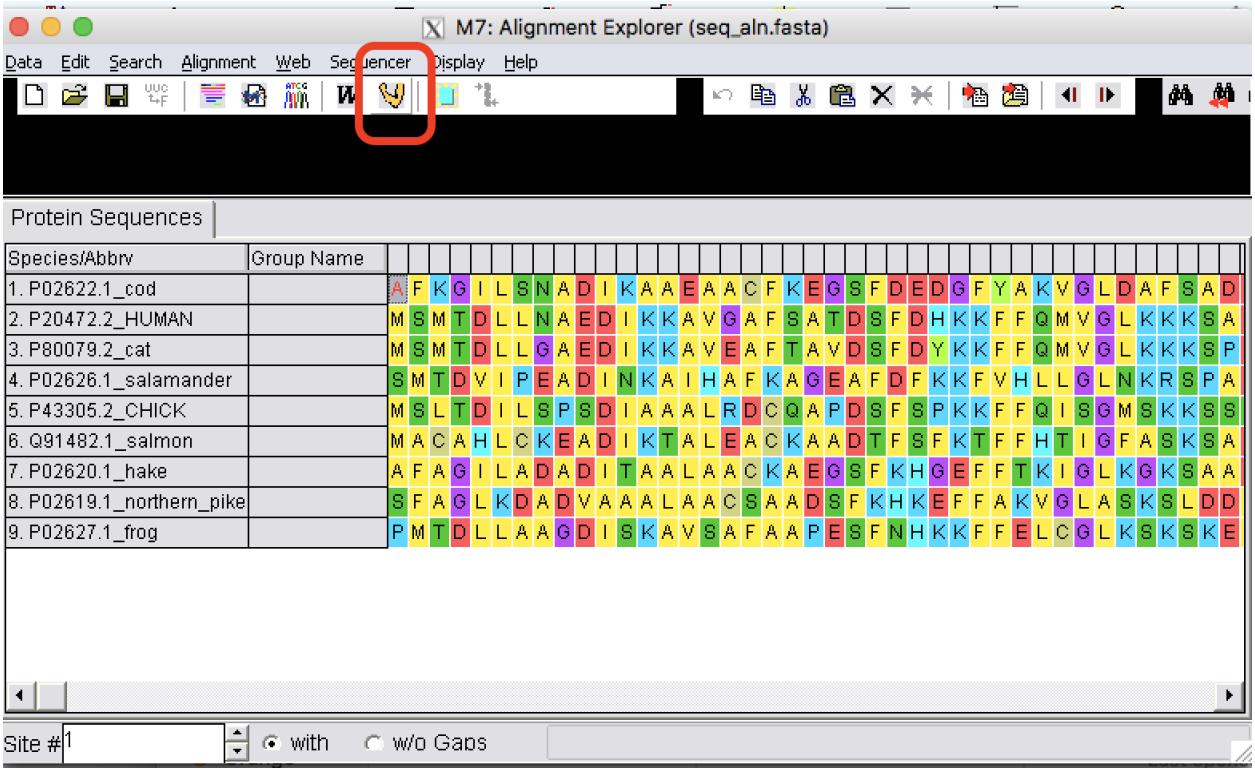

MEGA 7.0.21(7161111-i386)
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Clocks
Diagnose
Text
Help Docs
Examples
Citation
Report a Bug
Updates?
MEGA Links
Toolbar
Preferences
MEGA release #7161111-i386


MEGA 7.0.21(7161111-i386)
File
Analysis
Help
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Clocks
Diagnose
Edit/Build Alignment
Edit/View Sequencer Files (Trace)...
Open Saved Alignment Session...
Show Web Browser
Query Databanks
Do BLAST Search
Help Docs
Examples
Citation
Report a Bug
Updates?
MEGA Links
Toolbar
Preferences
MEGA release #7161111-i386


MEGA 7.0.21(7161111-i386)
File Analysis Help
Align Data Models Distance Diversity Phylogeny User Tree Ancestors Selection Rates Clocks Diagnose
M7: Alignment Editor
Select an Option
Create a new alignment
Open a saved alignment session
Retrieve sequences from a file
OK
Cancel
Help Docs Examples Citation Report a Bug Updates? MEGA Links Toolbar Preferences
MEGA release #7161111-i386


MEGA 7.0.21(7161111-i386)
File Analysis Help
Align Data Models Distance Diversity Phylogeny User Tree Ancestors Selection Rates Clocks Diagnose
Datatype for alignment
Are you building a DNA or Protein sequence alignment?
DNA
Protein
Cancel
Help Docs Examples Citation Report a Bug Updates? MEGA Links Toolbar Preferences
MEGA release #7161111-i386

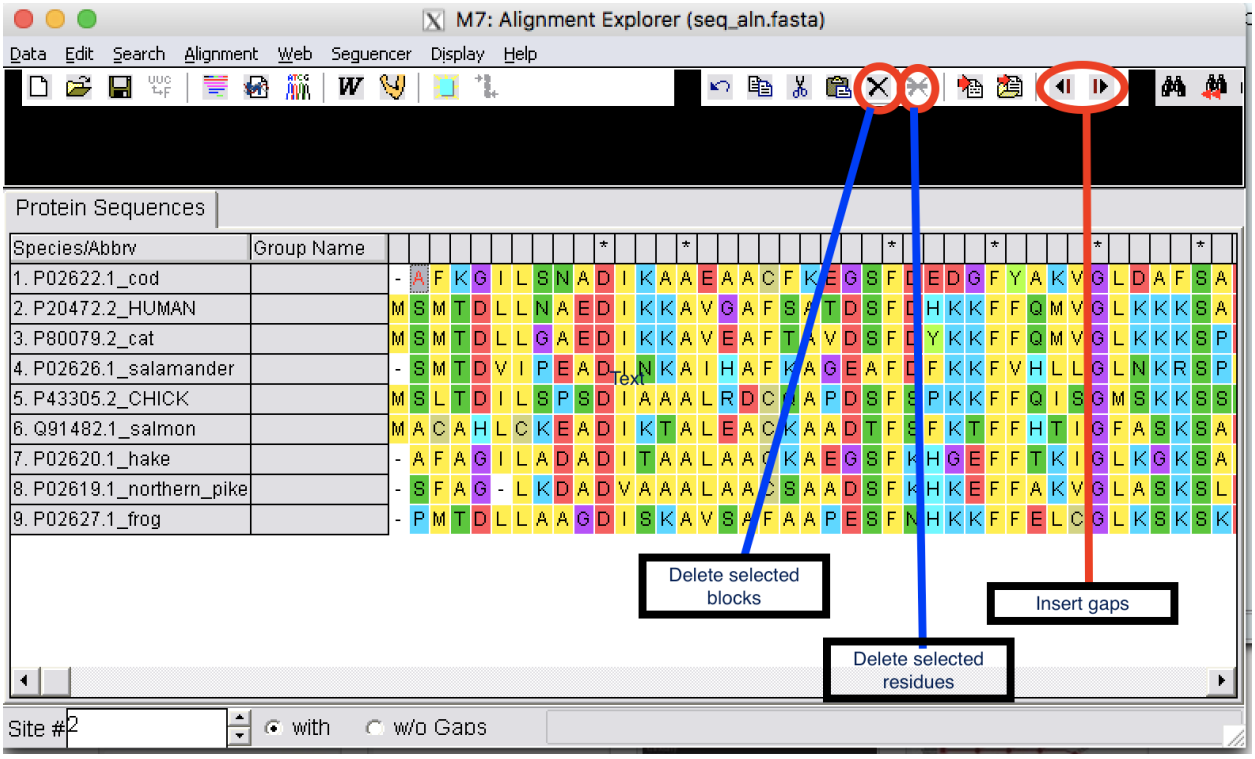

M7: Alignment Explorer
Data Edit Search Alignment Web Sequencer Display Help
Protein Sequences
Look in:
Inclass_activities_week5_2
seq_aln.fasta
File name:
Open
Files of type:
Supported sequence files
Cancel
Site #
with
w/o Gaps


M7: Alignment Explorer (seq_aln.fasta)
Data Edit Search Alignment Web Sequencer Display Help
Protein Sequences
Species/Abbrv
Group Name
1. P02622.1_cod
2. P20472.2_HUMAN
3. P80079.2_cat
4. P02626.1_salamander
5. P43305.2_CHICK
6. Q91482.1_salmon
7. P02620.1_hake
8. P02619.1_northern_pike
9. P02627.1_frog
Site #
with
w/o Gaps

1.1.2 Edit the alignments

Step 1 You can insert gaps, delete blocks, and delete residues.

Be very careful when you edit a sequence alignment. It should be biologically meaningful.



1.1.3 Exporting MSA

Mega allows to export the MSA in different formats.

CHAPTER 2

Steps of building a tree

2.1 Make multiple sequence alignment for Globin gene family

Step 1 Download globin.fasta from Blackboard and perform a MSA using MUSCLE (follow the steps we discussed last week).

Step 2 Examine the alignment to make sure it is correct and no additional editing is needed.

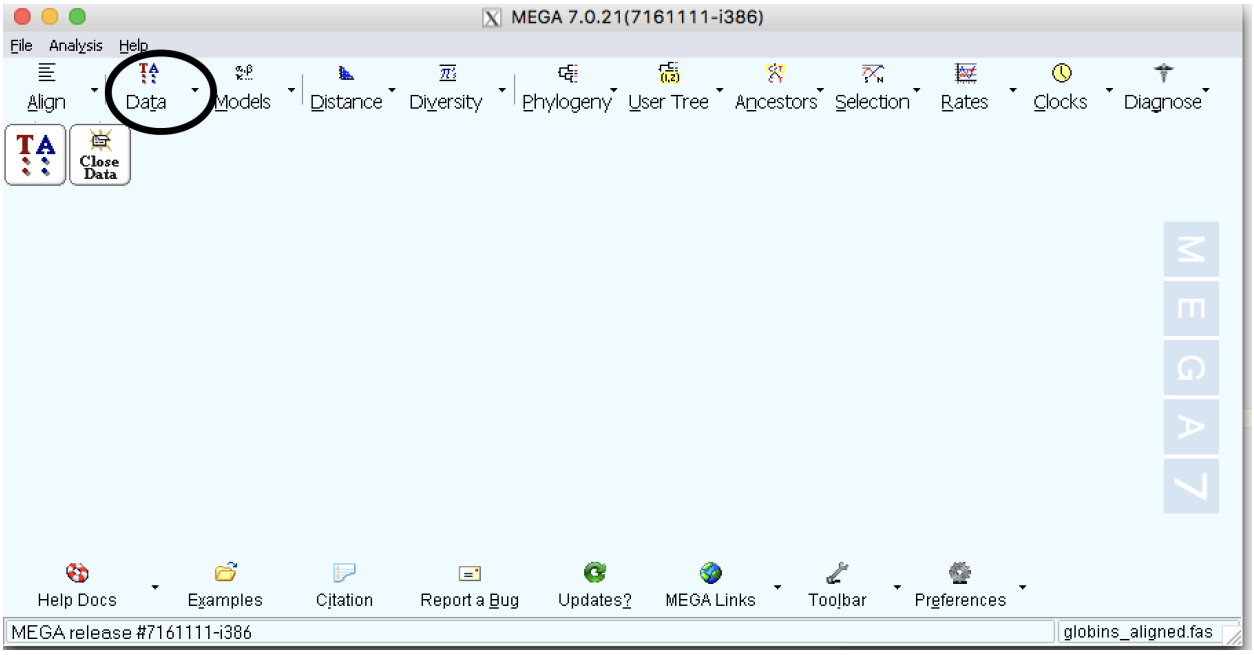
Step 2 Export the alignment as a fasta format file on your Desktop. Name it as globin_align

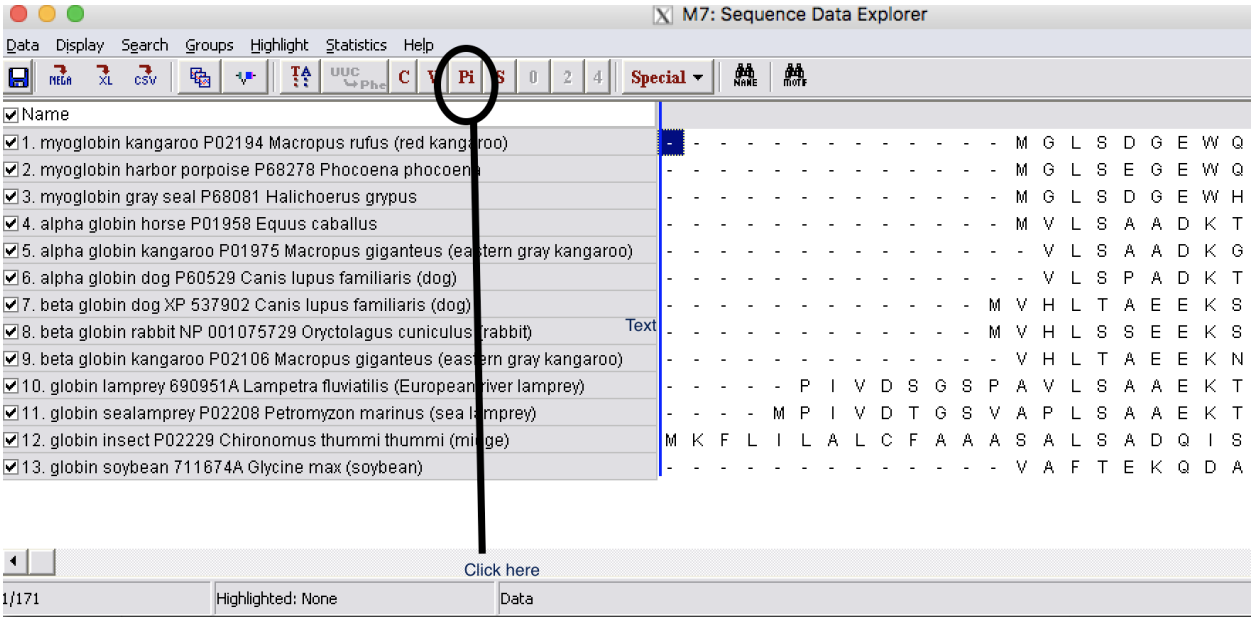
2.2 Find informative sites for Parsimony

Step 1 Open the alignment file you just created by going to using **Open file/Session** under **Data**

Step 2 Go to **Explore Active Data** under **Data**

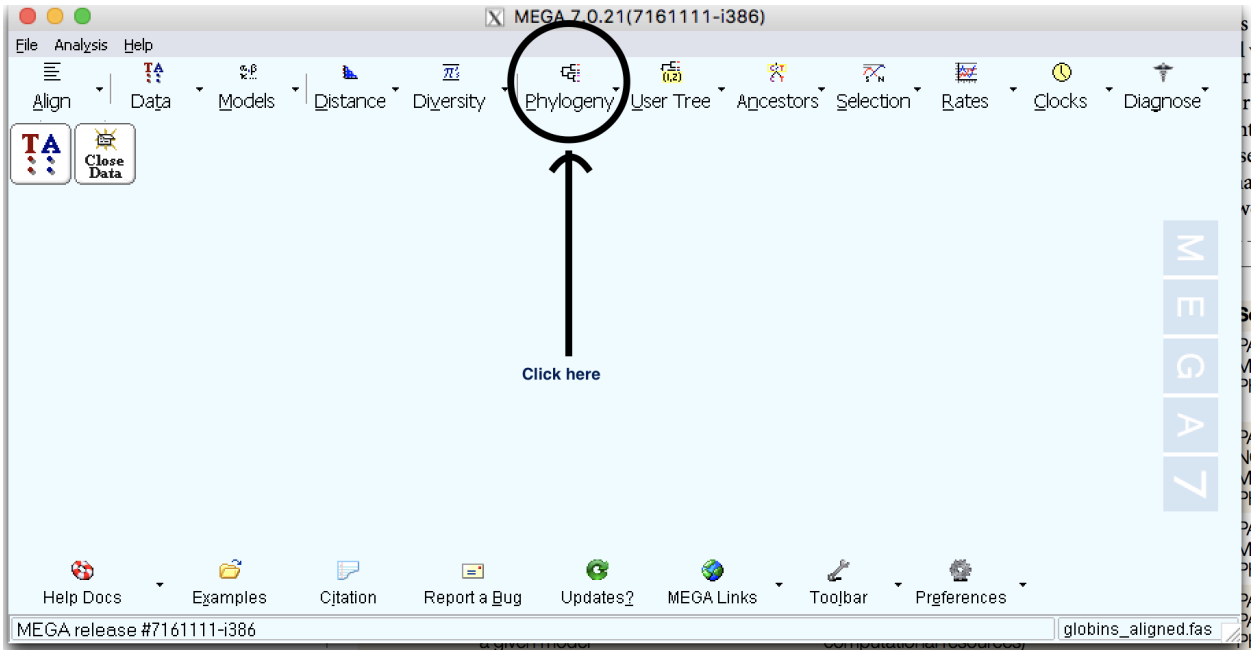
Step 3 Click on **Pi** button and this show site that are informative for Parsimony


MEGA 7.0.21(7161111-i386)
File Analysis Help
Align
Data
Models
Distance
Diversity
Phylogeny
User Tree
Ancestors
Selection
Rates
Clocks
Diagnose
Close Data
Help Docs
Examples
Citation
Report a Bug
Updates?
MEGA Links
Toolbar
Preferences
MEGA release #7161111-i386
globins_aligned.fas


M7: Sequence Data Explorer
Data Display Search Groups Highlight Statistics Help
Special
Name
1. myoglobin kangaroo P02194 Macropus rufus (red kangaroo)
2. myoglobin harbor porpoise P68278 Phocoena phocoena
3. myoglobin gray seal P68081 Halichoerus grypus
4. alpha globin horse P01958 Equus caballus
5. alpha globin kangaroo P01975 Macropus giganteus (eastern gray kangaroo)
6. alpha globin dog P60529 Canis lupus familiaris (dog)
7. beta globin dog XP 537902 Canis lupus familiaris (dog)
8. beta globin rabbit NP 001075729 Oryctolagus cuniculus (rabbit)
9. beta globin kangaroo P02106 Macropus giganteus (eastern gray kangaroo)
10. globin lamprey 690951A Lampetra fluviatilis (European river lamprey)
11. globin sealamprey P02208 Petromyzon marinus (sea lamprey)
12. globin insect P02229 Chironomus thummi thummi (midge)
13. globin soybean 711674A Glycine max (soybean)
Text
Click here
1/171
Highlighted: None
Data

2.3 Building Phylogenetic trees

Step 1 Click on **Phylogeny**



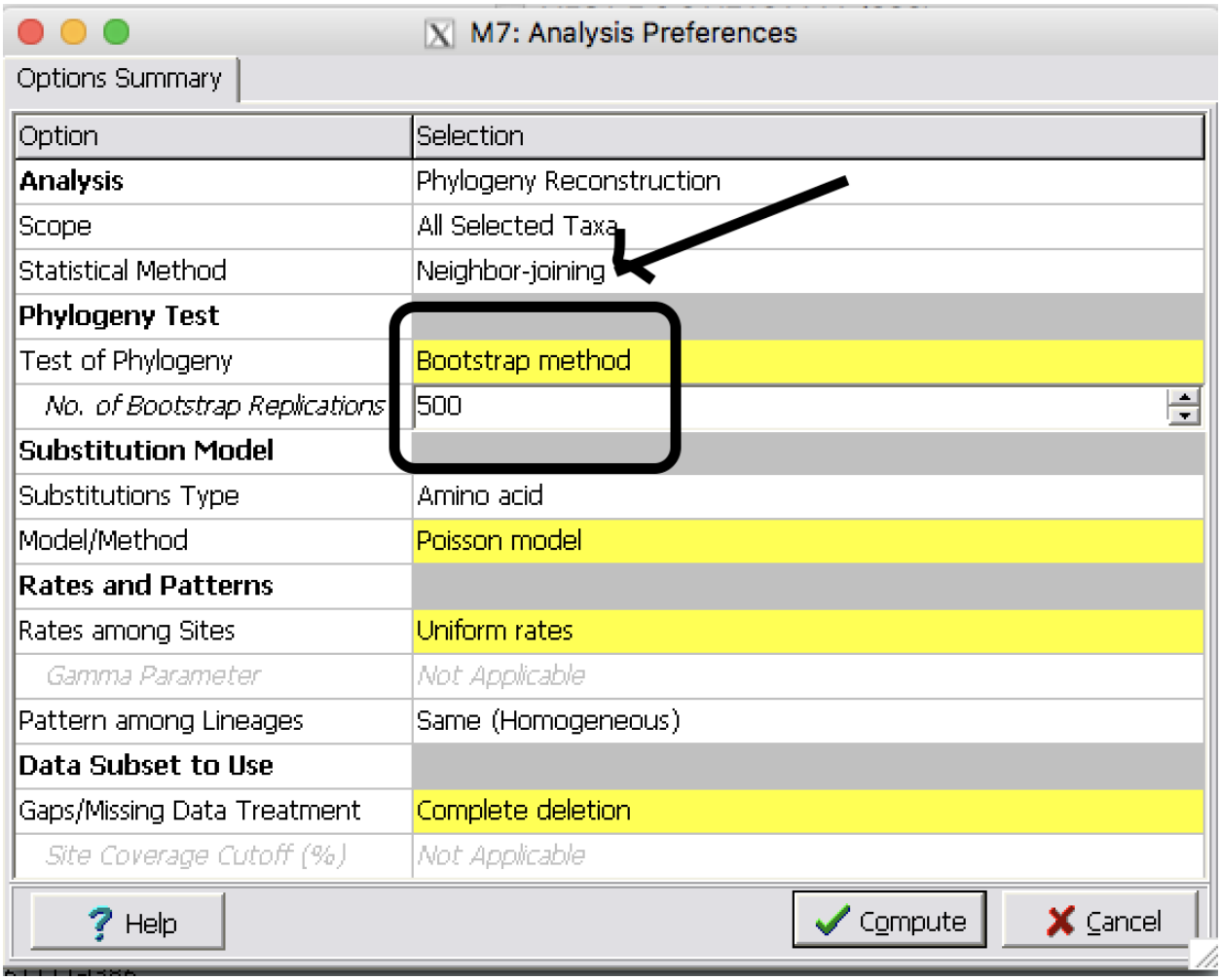
Step 2 Make Neighbor-Joining tree with **Bootstrap 500 replicates**

- **A** What relationships can you see in the tree?
- **B** What can you say about the statistical support for each relationship?
- **C** What should be the out-group?

Step 3 Save the tree as a pdf file by clicking on **Image** button

Step 4 Build a tree using Parsimony method with **50 Bootstrap** replicates (500 will be very slow).

- **A** What relationships can you see in the tree?
- **B** What can you say about the statistical support for each relationship?
- **C** Do you see the same relationships that you saw with NJ tree?


M7: Analysis Preferences
Options Summary
Option
Selection
Analysis
Phylogeny Reconstruction
Scope
All Selected Taxa
Statistical Method
Neighbor-joining
Phylogeny Test
Test of Phylogeny
Bootstrap method
No. of Bootstrap Replications
500
Substitution Model
Substitutions Type
Amino acid
Model/Method
Poisson model
Rates and Patterns
Rates among Sites
Uniform rates
Gamma Parameter
Not Applicable
Pattern among Lineages
Same (Homogeneous)
Data Subset to Use
Gaps/Missing Data Treatment
Complete deletion
Site Coverage Cutoff (%)
Not Applicable
Help
Compute
Cancel

CHAPTER 3

Steps of building a tree (Part II)

3.1 Make multiple sequence alignment for Globin gene family

Step 1 Download globin.fasta from Blackboard and perform a MSA using MUSCLE (follow the steps we discussed last week).

Step 2 Examine the alignment to make sure it is correct and no additional editing is needed.

Step 3 Export the alignment as a fasta format file on your Desktop. Name it as globin_align

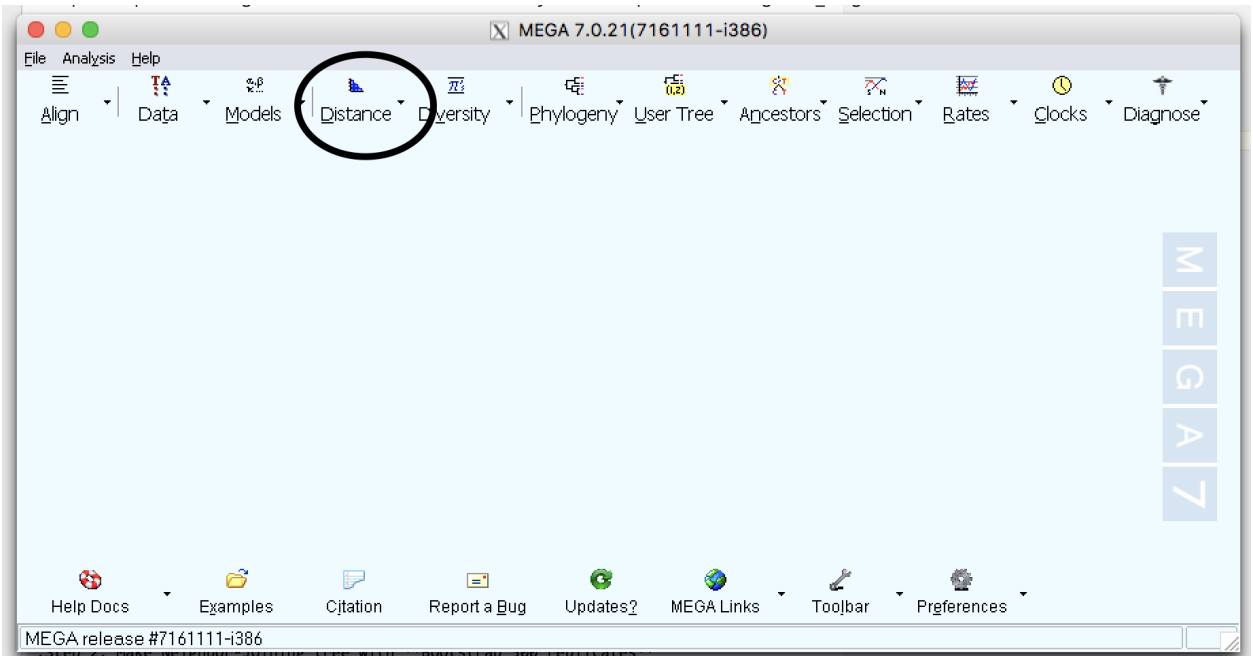
3.2 Find the best substitution model

Step 1 Calculate the distance using different substitution models :a: Select **Distance** and then **Compute Pairwise distance** :b: Calculate distance using the following methods

i No. of Differences

ii p-distances

iii Poisson model

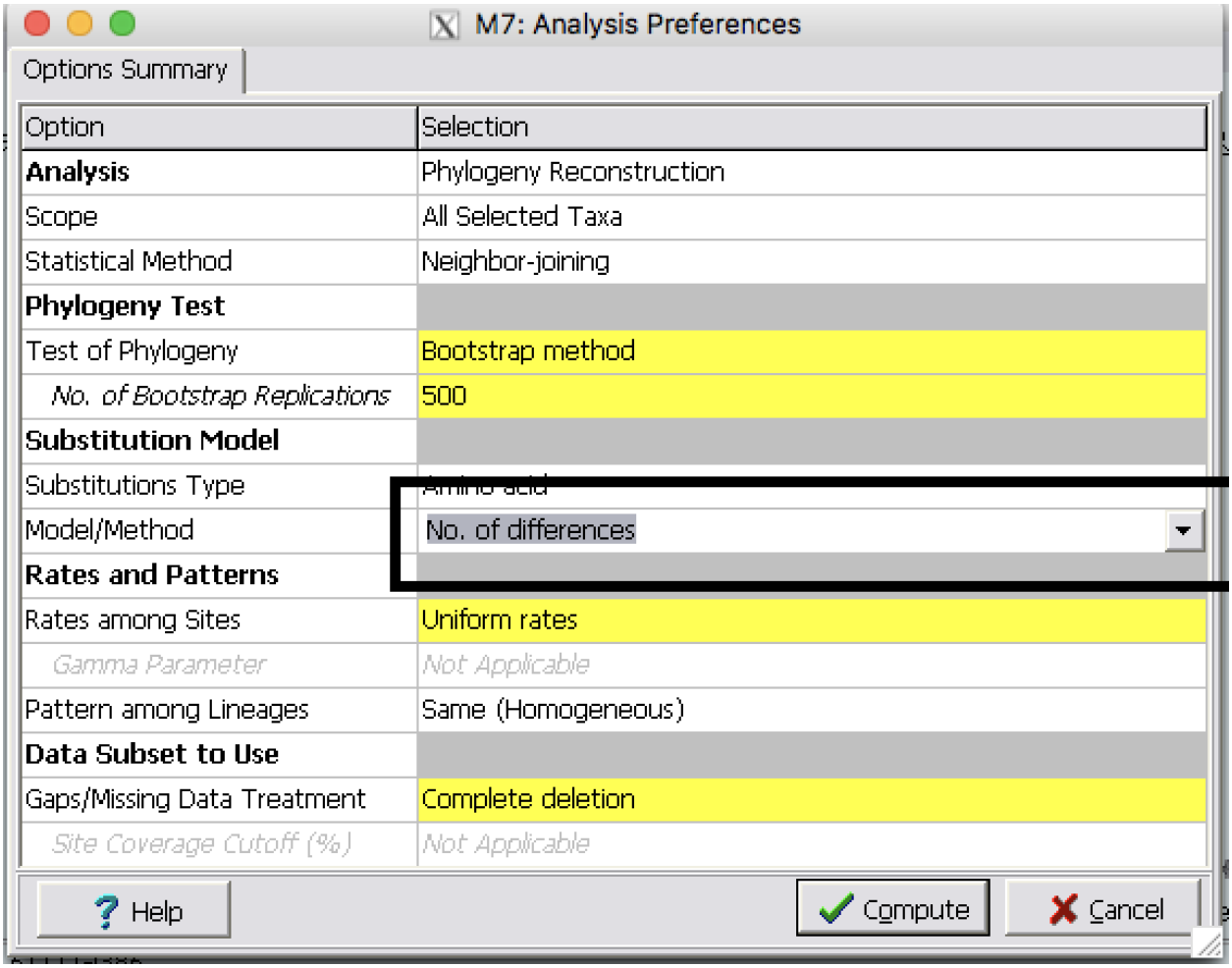


Step 2 Use the same alignment file and build three NJ trees using different substitution models:

a No. of Differences

b p-distances

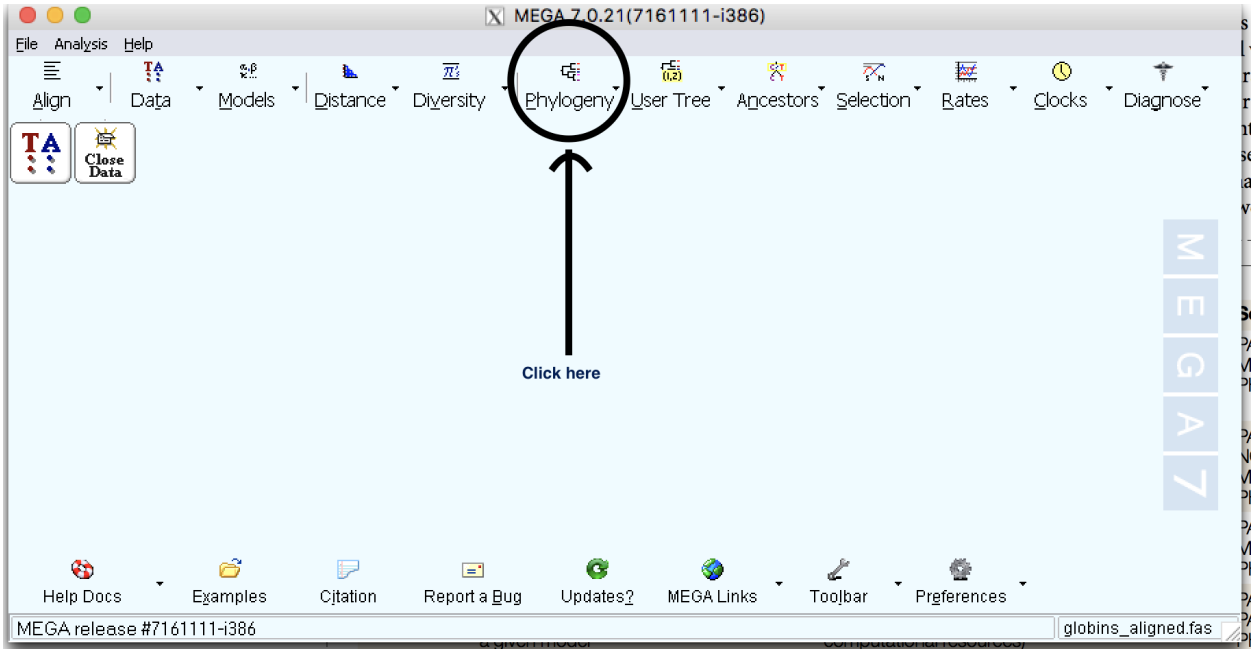
c Poisson model



Step 3 Best model based on ProtTest

3.3 Building Phylogenetic trees

Step 1 Click on **Phylogeny**



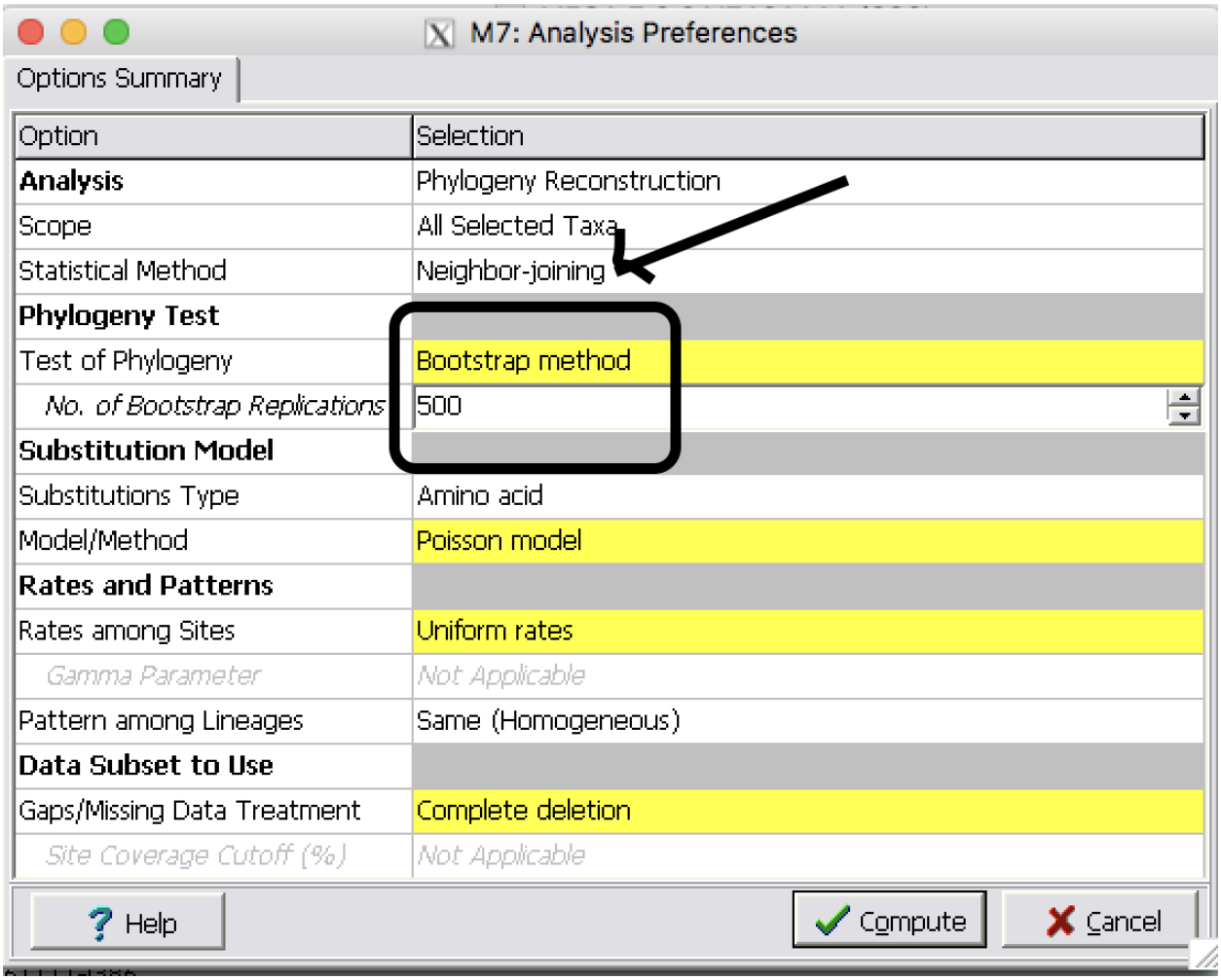
Step 2 Make Neighbor-Joining tree with **Bootstrap 500 replicates**

A What relationships can you see in the tree?

B What can you say about the statistical support for each relationship?

C What should be the out-group?

Step 3 Save the tree as a pdf file by clicking on **Image** button


M7: Analysis Preferences
Options Summary
Option
Selection
Analysis
Phylogeny Reconstruction
Scope
All Selected Taxa
Statistical Method
Neighbor-joining
Phylogeny Test
Test of Phylogeny
Bootstrap method
No. of Bootstrap Replications
500
Substitution Model
Substitutions Type
Amino acid
Model/Method
Poisson model
Rates and Patterns
Rates among Sites
Uniform rates
Gamma Parameter
Not Applicable
Pattern among Lineages
Same (Homogeneous)
Data Subset to Use
Gaps/Missing Data Treatment
Complete deletion
Site Coverage Cutoff (%)
Not Applicable
Help
Compute
Cancel

Step 4 Build a tree using Parsimony method with **50 Bootstrap** replicates (500 will be very slow).

A What relationships can you see in the tree?

B What can you say about the statistical support for each relationship?

C Do you see the same relationships that you saw with NJ tree?

Step 5 Bayesian inference of phylogeny

Follow this link to MrBayes online server

A Use the same alignment file

B In MrBayes select Poisson amino acid model with equal rates of substitution.

C Select prior parameters (e.g. equal, fixed frequencies for the states; equal probability for all topologies; unconstrained branch lengths).

D Run 1,000,000 trials for Monte Carlo Markov Chain estimation of the posterior distribution.

E Obtain phylogram

F Export tree files

G View in MEGA software

Input:

Alignment

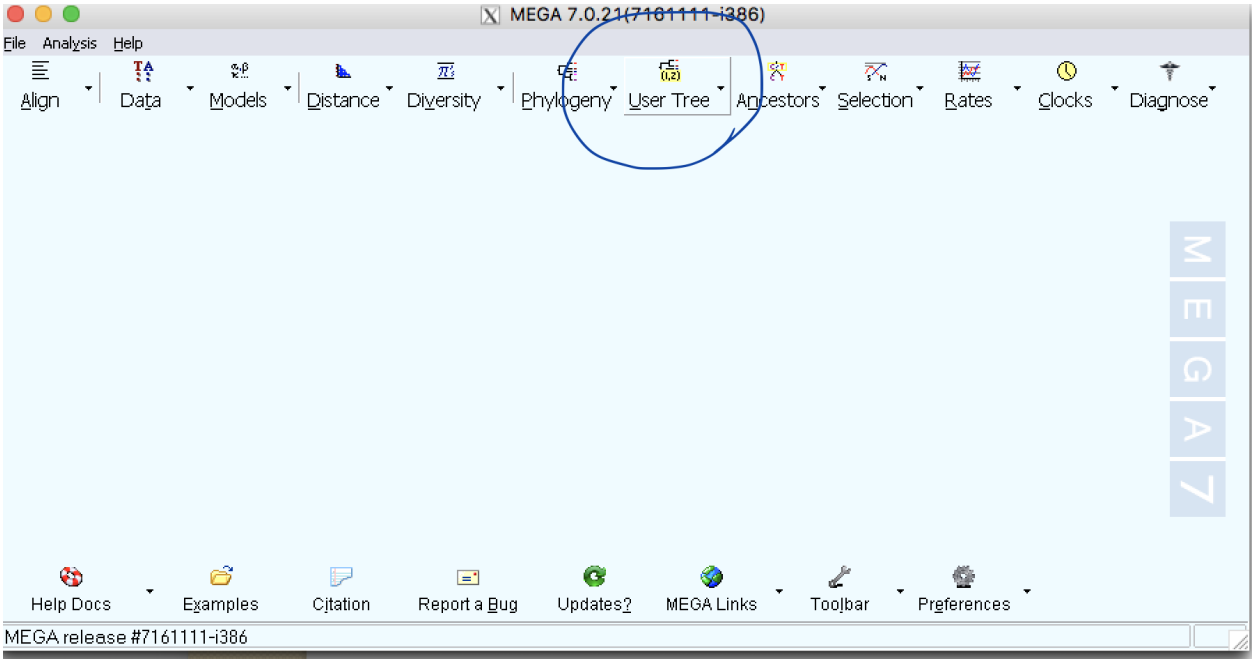
Outputs:

Tree in Newick format (automatically recognized by MEGA if installed)

MrBayes logs

Taxon names association table

Download taxon names association table



CHAPTER 4

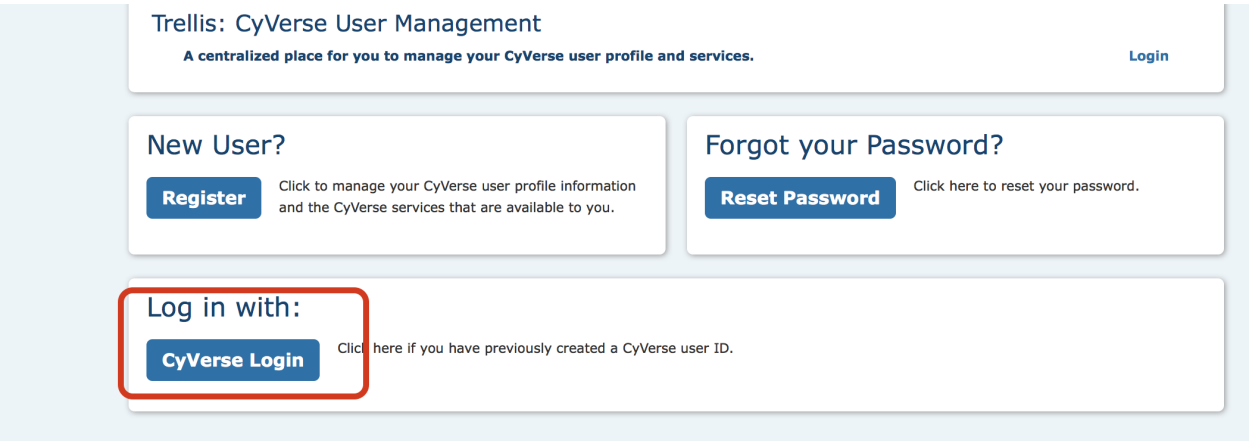
FastQC analysis using Cyverse Discovery Environment (DE)

Data we are using for this analysis came from Loraine et al, 2015 study. In the original study, there are 10 samples (Five Controls and heat treated). Here we are using only 3 samples for each group (3 control and 3 heat treated). These files were downloaded from NCBI's Short Read Archive (SRA) using SRA toolkit.

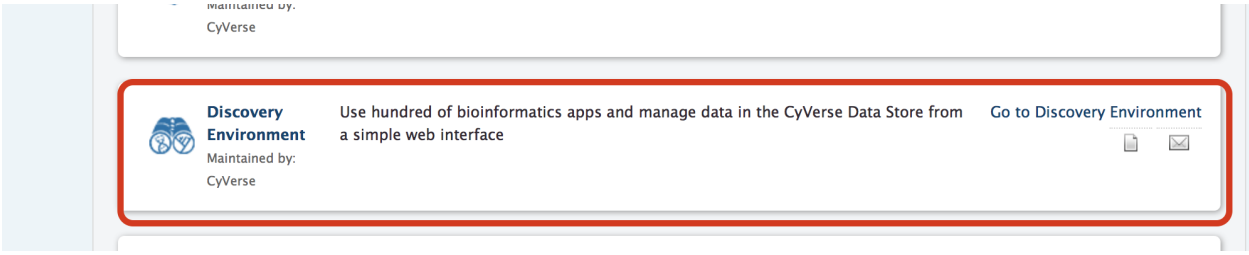
First step of the data analysis is to check the quality of the sequences. For this purpose, we are using the FastQC tool on Cyverse DE.

4.1 Step 1: Login into Cyverse DE

First login to your Cyverse account using your name and password.



Then, go to your DE account.

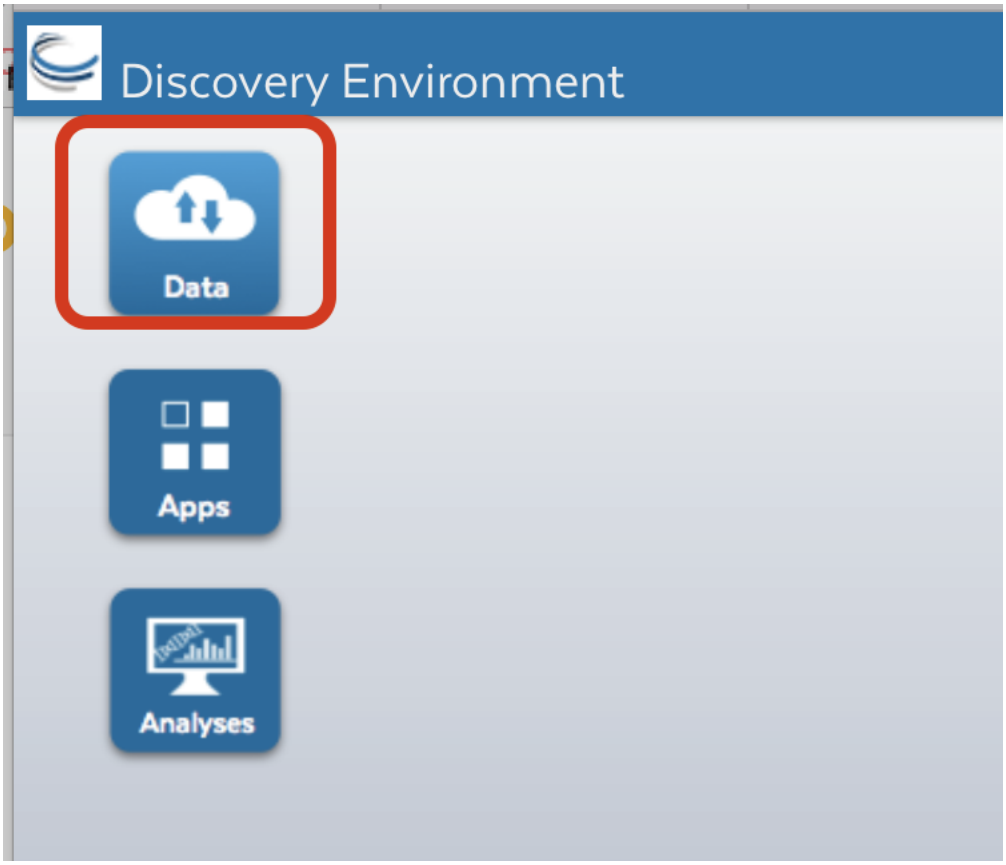


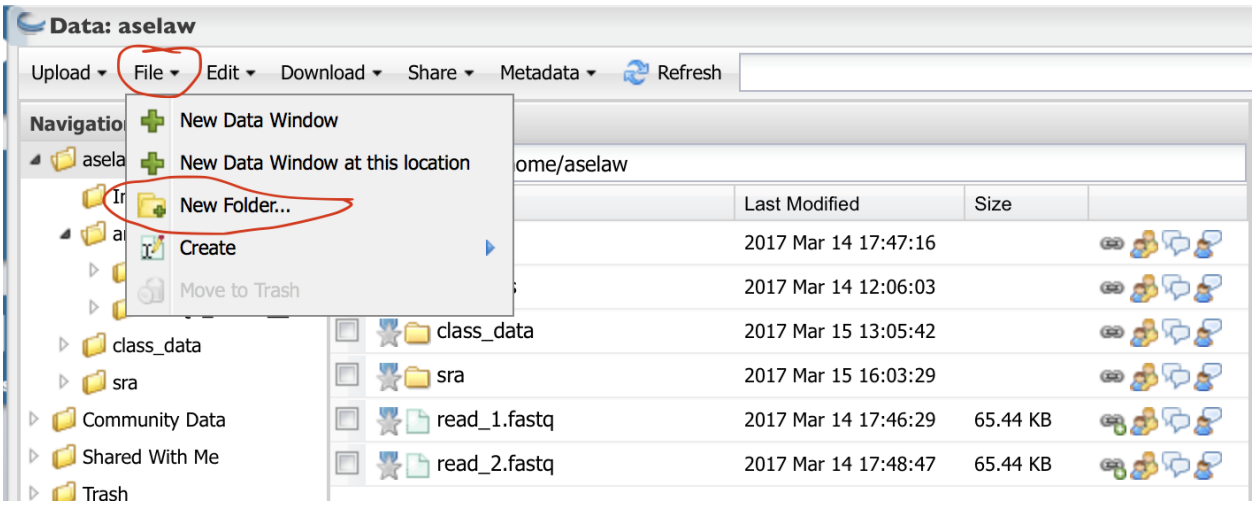
4.2 Step 2: Getting data into Cyverse Discovery Environment

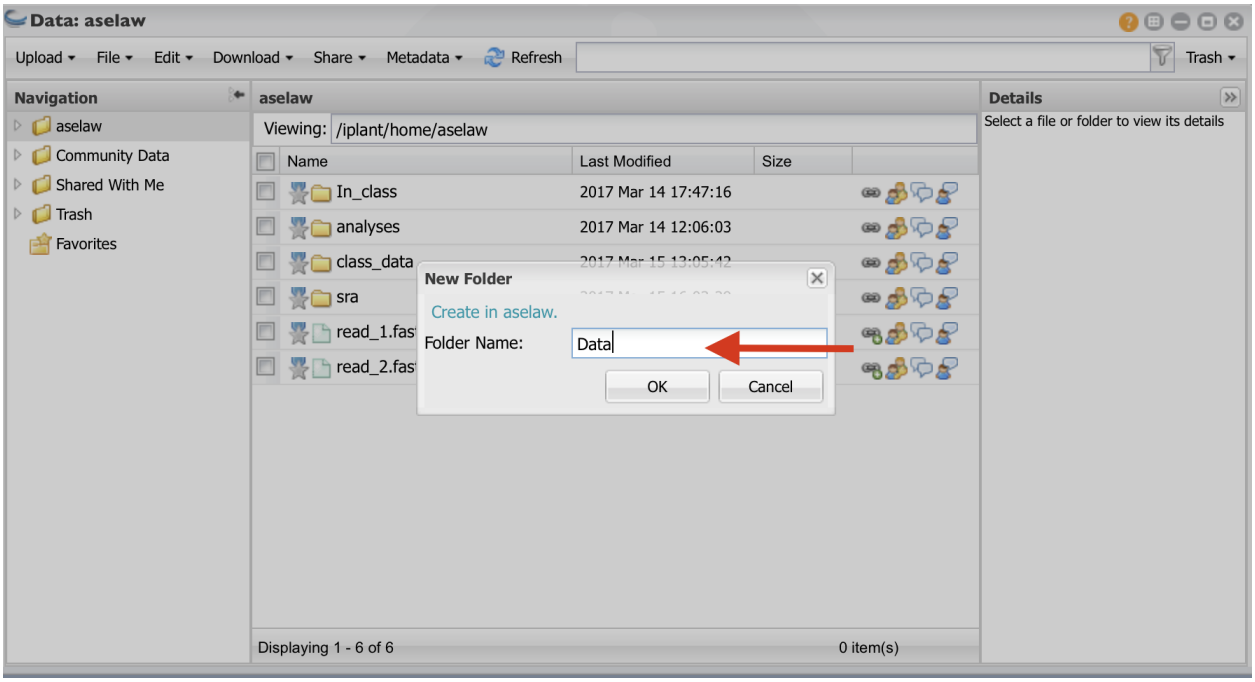
a. Click on "Data" button

b. Click on "File" and then "New Folder"

c. Create a folder called "Data" and click "OK". **Create another folder called "Analysis".**


Discovery Environment
Data
Apps
Analyses


Data: aselaw
Upload
File
Edit
Download
Share
Metadata
Refresh
New Data Window
New Data Window at this location
New Folder...
Create
Move to Trash
Last Modified
Size
2017 Mar 14 17:47:16
2017 Mar 14 12:06:03
class_data
2017 Mar 15 13:05:42
sra
2017 Mar 15 16:03:29
read_1.fastq
2017 Mar 14 17:46:29
65.44 KB
read_2.fastq
2017 Mar 14 17:48:47
65.44 KB
Community Data
Shared With Me
Trash



d. Click on the "Data" folder to enter into it. Click on "Upload" and then "Import from URL"

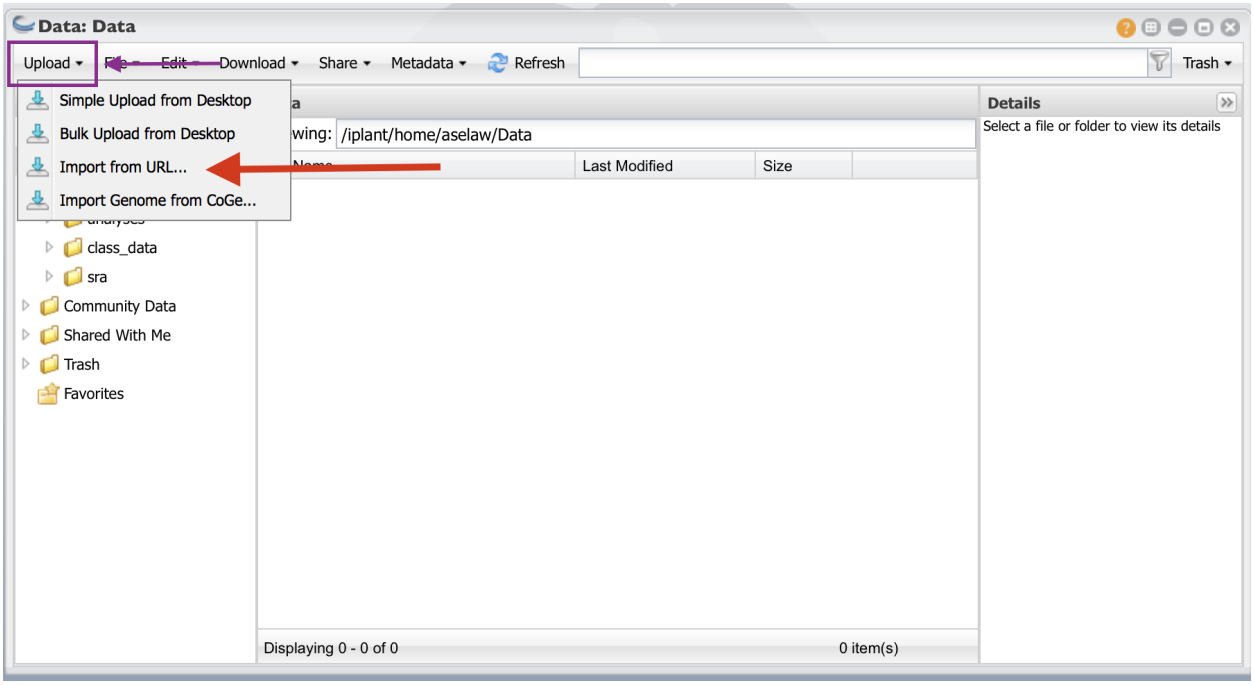
e. I have create public links for fastq files. Copy and paste URLs in the box (one for each box). **You will need to do this for all 12 URLs.** Then click on "Import from URL"

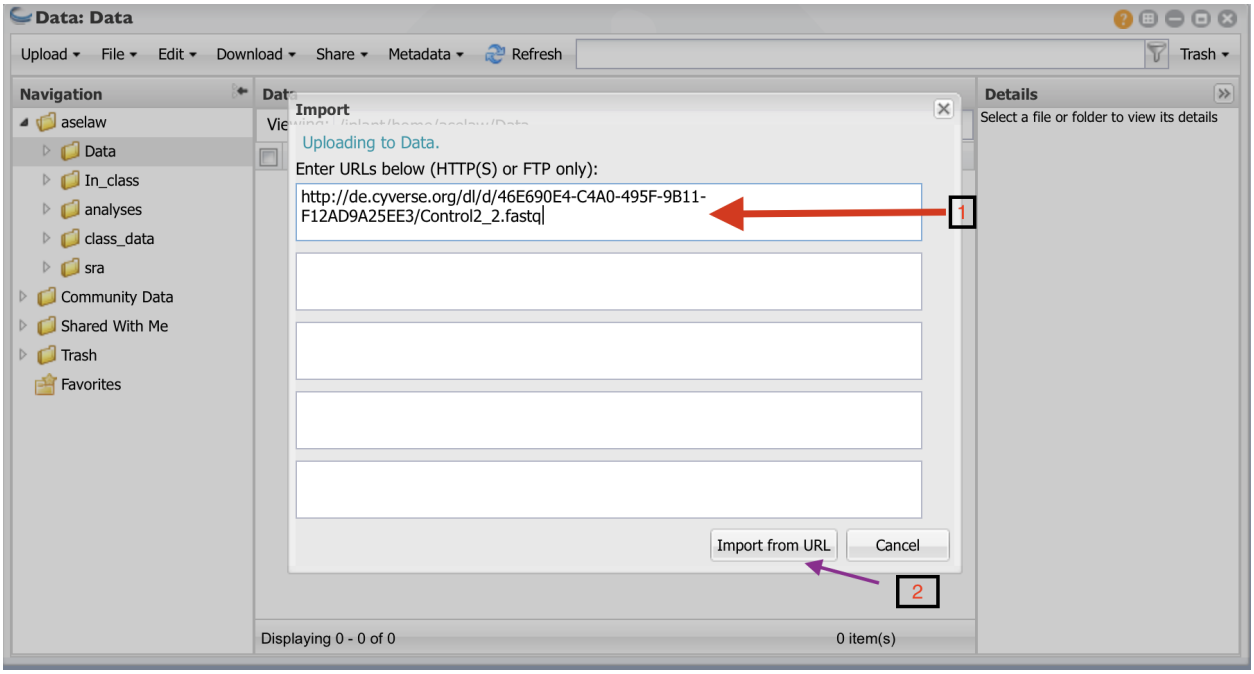
4.2.1 URLs

http://de.cyverse.org/dl/d/5B50EFE6-D0BA-4833-980E-E81E5B63C15E/Control1_1.fastq

http://de.cyverse.org/dl/d/BBFB60AC-8855-40AC-9634-7C62F5B9B02D/Control1_2.fastq

http://de.cyverse.org/dl/d/2AB5824F-73BA-4C6B-8530-457609F632BA/Control2_1.fastq


Data: Data
Upload
File
Edit
Download
Share
Metadata
Refresh
Trash
Simple Upload from Desktop
Bulk Upload from Desktop
Import from URL...
Import Genome from CoGe...
/iplant/home/aselaw/Data
Last Modified
Size
class_data
sra
Community Data
Shared With Me
Trash
Favorites
Details
Select a file or folder to view its details
Displaying 0 - 0 of 0
0 item(s)


Data: Data
Upload
File
Edit
Download
Share
Metadata
Refresh
Trash
Navigation
aselaw
Data
In_class
analyses
class_data
sra
Community Data
Shared With Me
Trash
Favorites
Import
Uploading to Data.
Enter URLs below (HTTP(S) or FTP only):
http://de.cyverse.org/dl/d/46E690E4-C4A0-495F-9B11-F12AD9A25EE3/Control2_2.fastq
1
Import from URL
Cancel
2
Details
Select a file or folder to view its details
Displaying 0 - 0 of 0
0 item(s)

http://de.cyverse.org/dl/d/46E690E4-C4A0-495F-9B11-F12AD9A25EE3/Control2_2.fastq

http://de.cyverse.org/dl/d/7FEE6359-24AE-478D-A0B1-C6D2CA09E45E/Control3_1.fastq

http://de.cyverse.org/dl/d/8FBB264D-F0CA-4F2C-821A-DB1C709315B2/Control3_2.fastq

http://de.cyverse.org/dl/d/E7AD135C-F2BC-445C-BBC2-695B1D76B010/Heat1_1.fastq

http://de.cyverse.org/dl/d/46093383-493A-4D4E-A607-D3E56916DF59/Heat1_2.fastq

http://de.cyverse.org/dl/d/9668B243-7009-4AD3-BBDA-350D6A60119D/Heat2_1.fastq

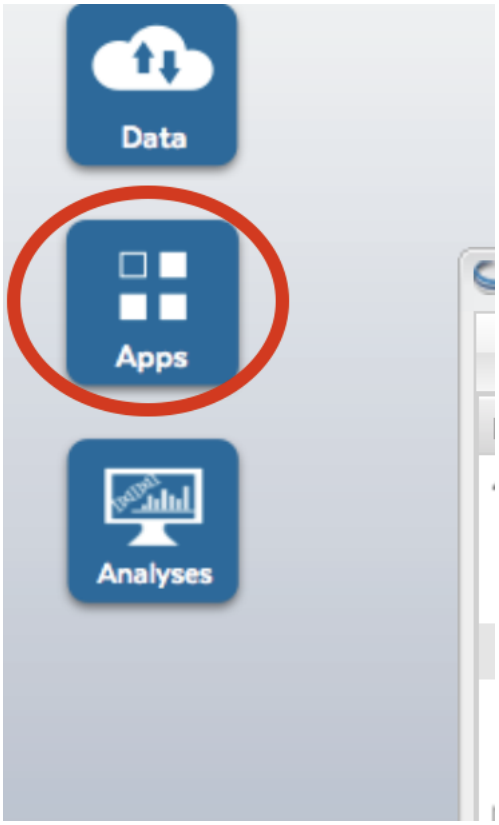
http://de.cyverse.org/dl/d/FE1C3CC3-9133-4244-BCBB-816B8D2D5F97/Heat2_2.fastq

http://de.cyverse.org/dl/d/D635B6EE-BE26-4BC4-A058-3E51B1AA69C4/Heat3_1.fastq

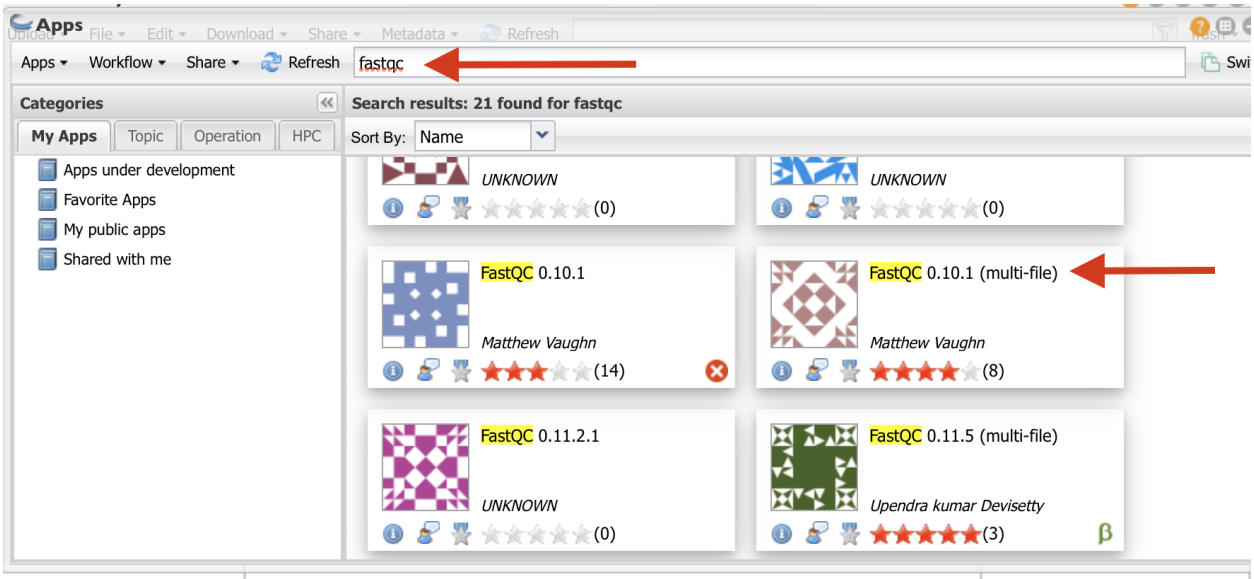
http://de.cyverse.org/dl/d/F88561AF-CFF2-4FC8-B6B4-D8623779BB24/Heat3_2.fastq

4.3 Step 3: Performing FastQC analysis:

a. Click on “Apps” button.



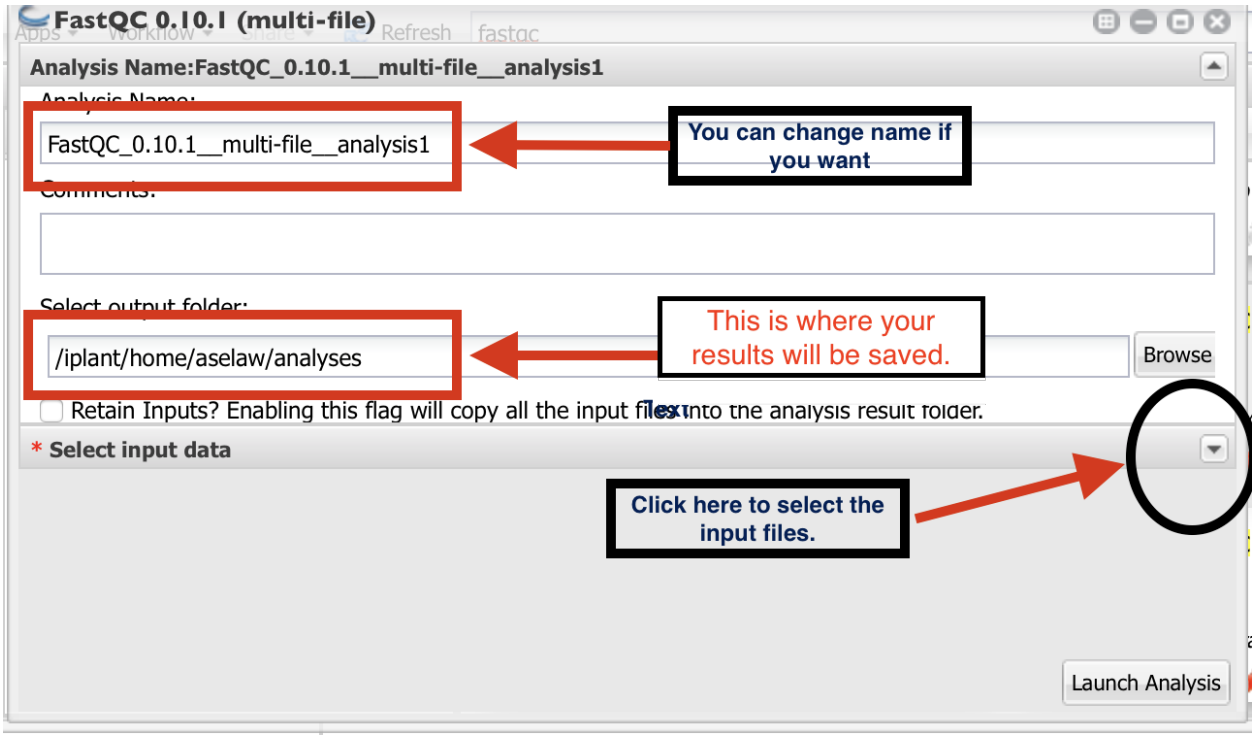
b. Type "fastqc" in the search window and select the app shown in red arrow.



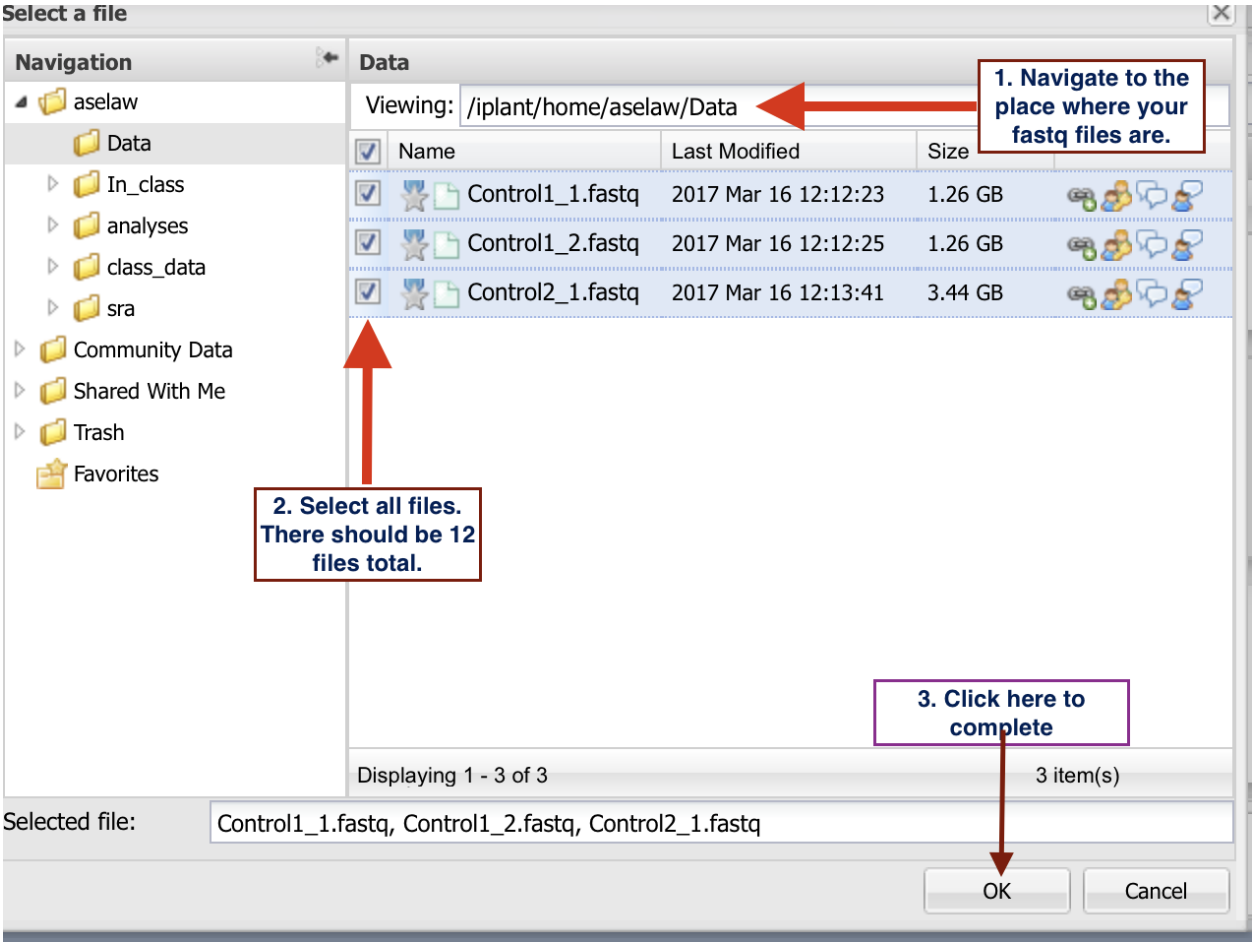
c. **Follow the direction as in the figure to select the folder where your results will be saved.** Then, click on the small downward arrow (black circle).

d. Click on "+" sign to select the fastq files.

e. **Go to the folder where you have your fastq files and select them as indicated in the** figure below. Then launch the analysis. Once the analysis is complete, you will be notified via email.


FastQC 0.10.1 (multi-file)
Refresh
fastqc
Analysis Name:FastQC_0.10.1__multi-file__analysis1
FastQC_0.10.1__multi-file__analysis1
You can change name if you want
/iplant/home/aselaw/analyses
This is where your results will be saved.
Browse
Retain Inputs? Enabling this flag will copy all the input files into the analysis result folder.
* Select input data
Click here to select the input files.
Launch Analysis


FastQC 0.10.1 (multi-file)
Refresh
fastqc
Analysis Name:FastQC_0.10.1__multi-file__analysis1
* Select input data
* Select FASTQ file(s):
Click here to add files
Add
Delete
Name
Select multiple input files.
Tip: You can also drag and drop files from the Data window.
Launch Analysis


Select a file
Navigation
Data
aselaw
Data
In_class
analyses
class_data
sra
Community Data
Shared With Me
Trash
Favorites
Viewing: /iplant/home/aselaw/Data
1. Navigate to the place where your fastq files are.
Name
Last Modified
Size
Control1_1.fastq
2017 Mar 16 12:12:23
1.26 GB
Control1_2.fastq
2017 Mar 16 12:12:25
1.26 GB
Control2_1.fastq
2017 Mar 16 12:13:41
3.44 GB
2. Select all files. There should be 12 files total.
3. Click here to complete
Displaying 1 - 3 of 3
3 item(s)
Selected file:
Control1_1.fastq, Control1_2.fastq, Control2_1.fastq
OK
Cancel

4.4 Reference:

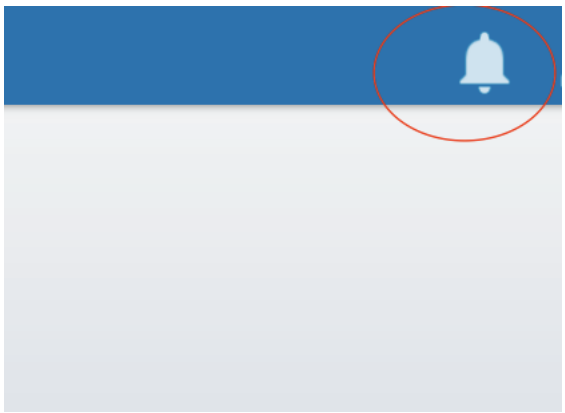
Loraine AE, Blakley IC, Jagadeesan S, Harper J, Miller G, Firon N. Analysis and Visualization of RNA-Seq Expression Data Using RStudio, Bioconductor, and Integrated Genome Browser. *Methods Mol Biol.* 2015;1284:481-501. doi: 10.1007/978-1-4939-2444-8_24. PubMed PMID: 25757788.

CHAPTER 5

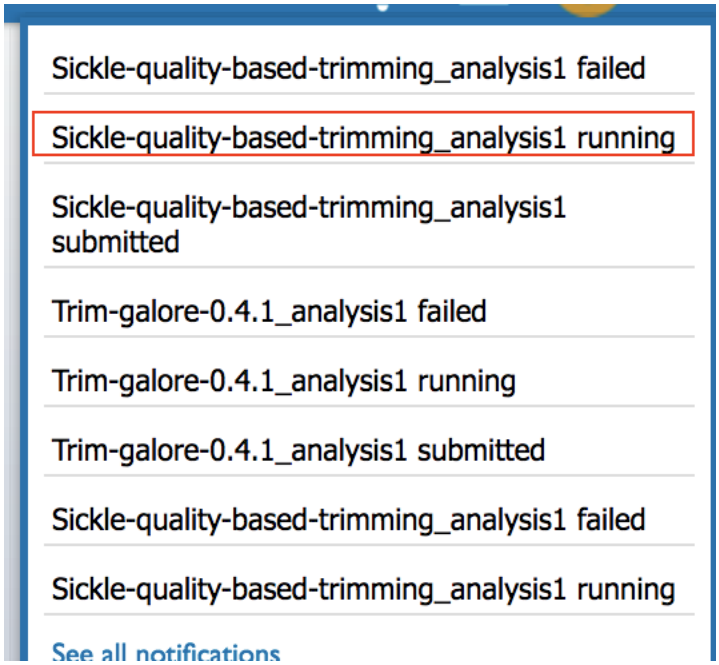
Relaunching a stalled analysis in Cyverse Discovery Environment

If your analysis is appeared to be stalled, you could try restarting it.

5.1 Step 1: Click on the message icon

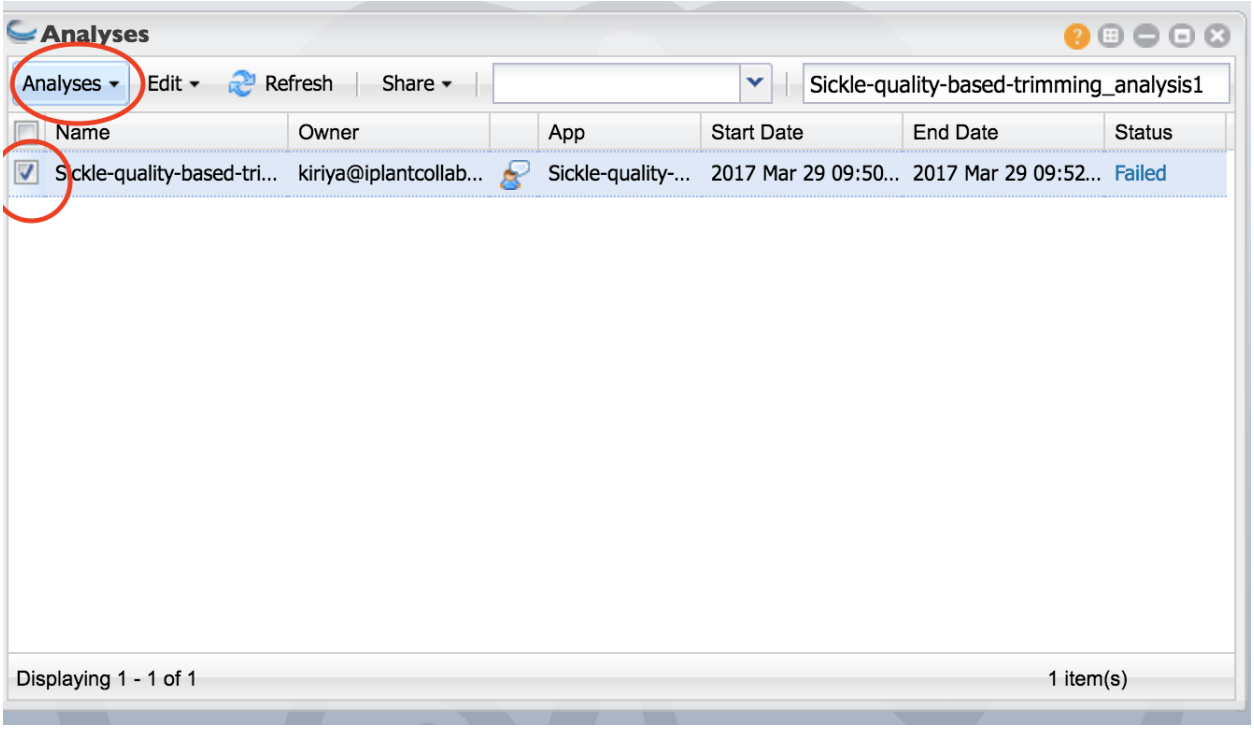


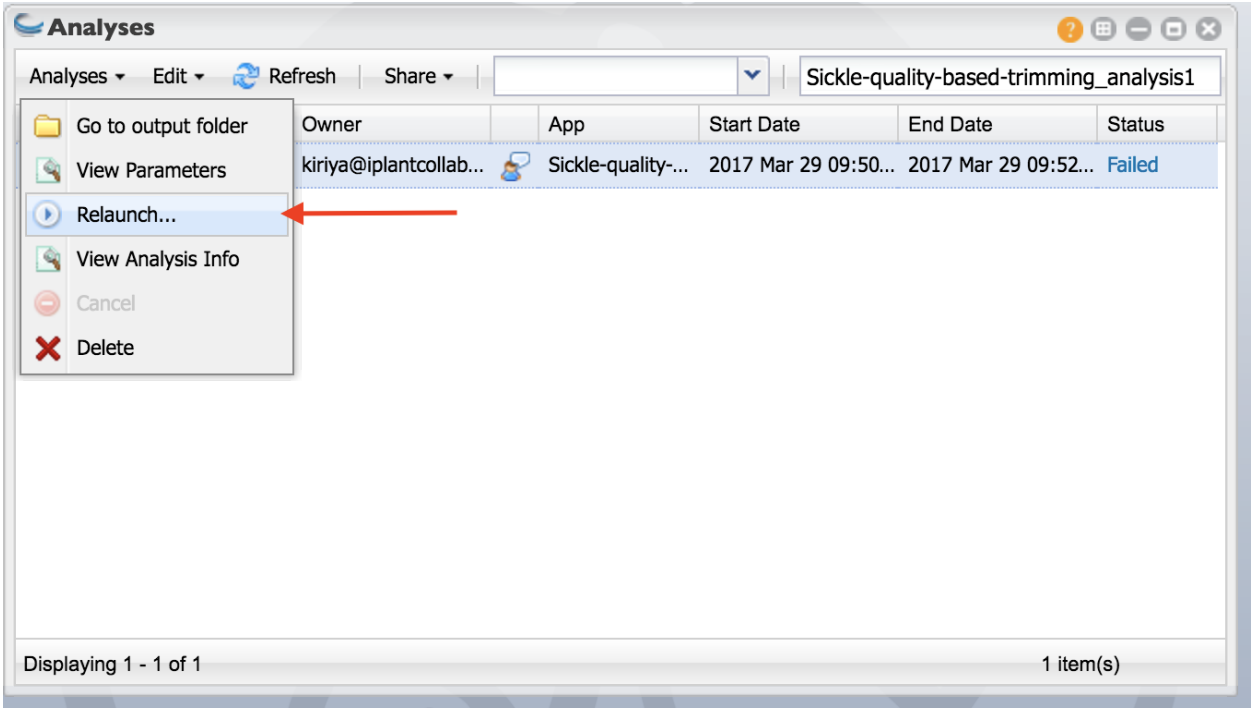
5.2 Step 2: Click on the analysis that appears to be stalled



5.3 Step 3: Check the small box and click on analysis

5.4 Step 4: Click on the relaunch button


Analyses
Analyses
Edit
Refresh
Share
Sickle-quality-based-trimming_analysis1
Name
Owner
App
Start Date
End Date
Status
Sickle-quality-based-tri...
kiriya@iplantcollab...
Sickle-quality-...
2017 Mar 29 09:50...
2017 Mar 29 09:52...
Failed
Displaying 1 - 1 of 1
1 item(s)


Analyses
Analyses
Edit
Refresh
Share
Sickle-quality-based-trimming_analysis1
Go to output folder
View Parameters
Relaunch...
View Analysis Info
Cancel
Delete
Owner
App
Start Date
End Date
Status
kiriya@iplantcollab...
Sickle-quality-...
2017 Mar 29 09:50...
2017 Mar 29 09:52...
Failed
Displaying 1 - 1 of 1
1 item(s)

Once the analysis window appears, launch the analysis.

CHAPTER 6

Adapter and quality trimming using trim-galore

We are going to use Trim-galore to trim adapters, and poor quality bases. This tool has several advantages. It allows selection multiple files. You can also select both forward and reverse reads. If you want to read more about Trim-galore, please visit their website. Also, Trim-galore is a wrapper for Cutadapt , which is the actual tool that performs the trimming.

Please follow the tutorial carefully.

6.1 Step 1: Launching Trim-galore

1. Click on *App*.

2. In the finder window type "trim-galore"

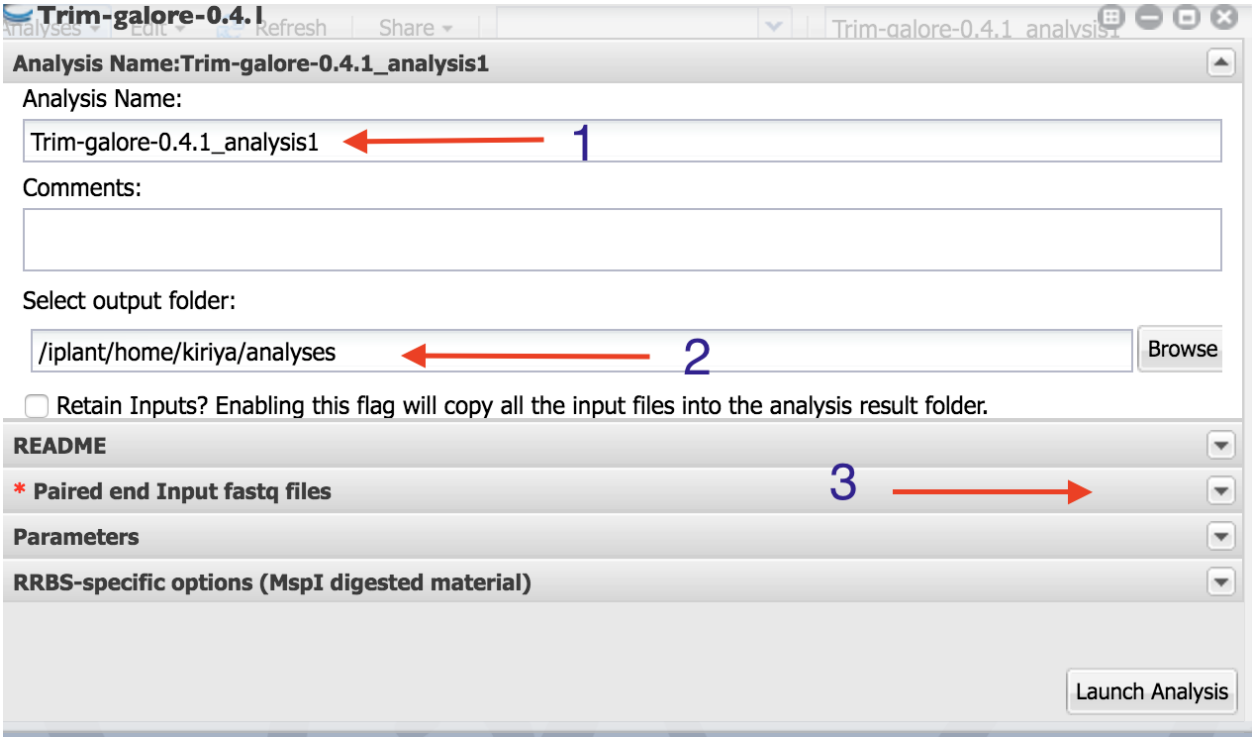
3. Select "trim-galore-0.4.1".

6.2 Step 2: Selecting output folder

As indicated in the figure: 1. Name your analysis as you want

2. Select the output folder where your analysis is going to be

3. Click on "Paired end Input fastq files"

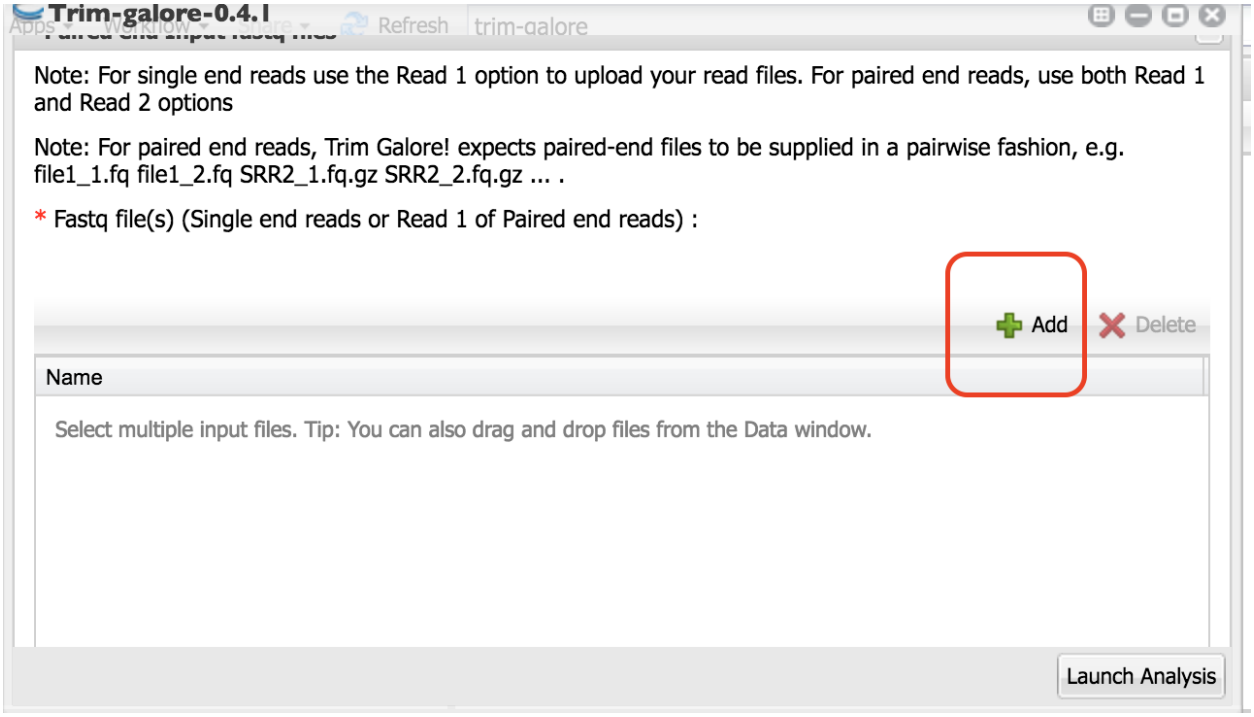


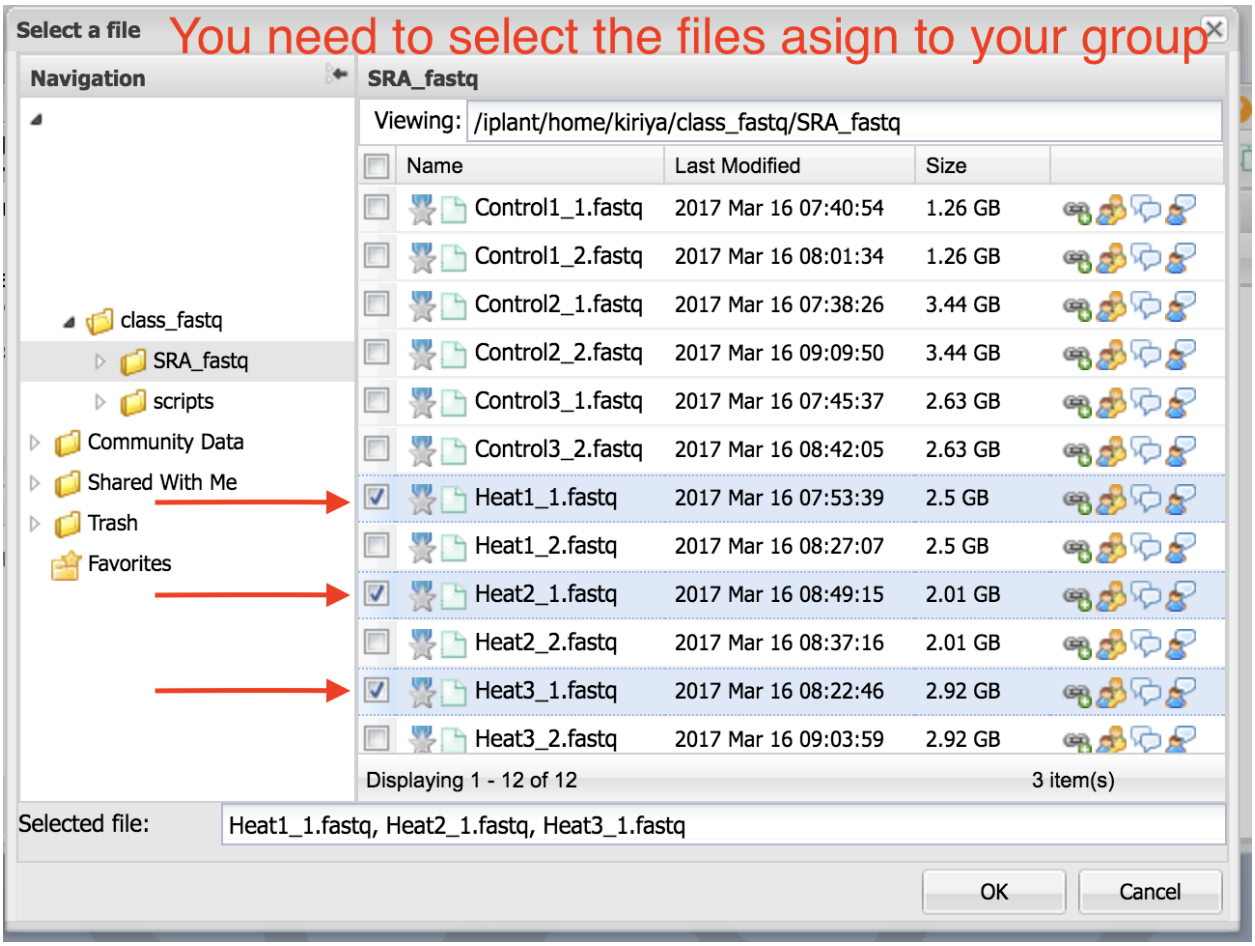
6.3 Step 3: Selecting input files

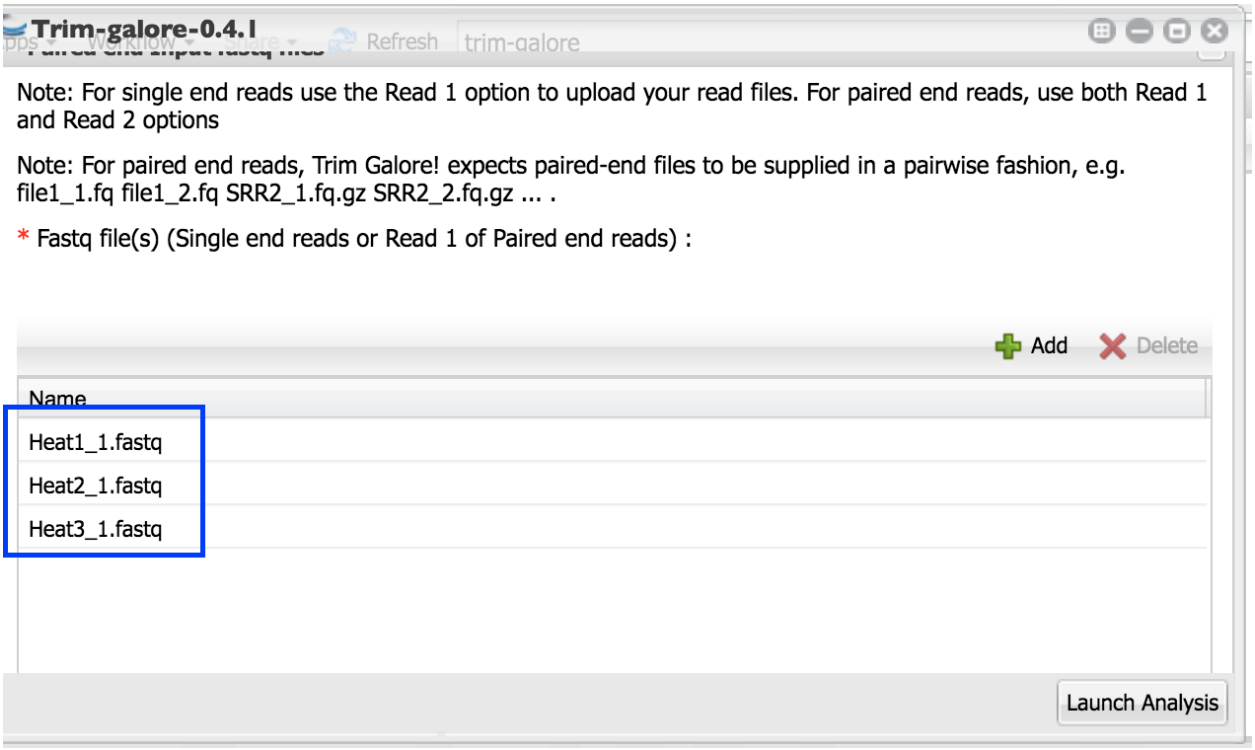
1. Click on the Green "+" sign.

2. Navigate to the folder where your samples are located. Select only the **first read files**. Click "OK".

3. You should all your first read files selected like this.


Trim-galore-0.4.1
Refresh
trim-galore
Note: For single end reads use the Read 1 option to upload your read files. For paired end reads, use both Read 1 and Read 2 options
Note: For paired end reads, Trim Galore! expects paired-end files to be supplied in a pairwise fashion, e.g. file1_1.fq file1_2.fq SRR2_1.fq.gz SRR2_2.fq.gz ... .
* Fastq file(s) (Single end reads or Read 1 of Paired end reads) :
Add
Delete
Name
Select multiple input files. Tip: You can also drag and drop files from the Data window.
Launch Analysis


Select a file
You need to select the files asign to your group
Navigation
SRA_fastq
Viewing: /iplant/home/kiriya/class_fastq/SRA_fastq
Name
Last Modified
Size
Control1_1.fastq 2017 Mar 16 07:40:54 1.26 GB
Control1_2.fastq 2017 Mar 16 08:01:34 1.26 GB
Control2_1.fastq 2017 Mar 16 07:38:26 3.44 GB
Control2_2.fastq 2017 Mar 16 09:09:50 3.44 GB
Control3_1.fastq 2017 Mar 16 07:45:37 2.63 GB
Control3_2.fastq 2017 Mar 16 08:42:05 2.63 GB
Heat1_1.fastq 2017 Mar 16 07:53:39 2.5 GB
Heat1_2.fastq 2017 Mar 16 08:27:07 2.5 GB
Heat2_1.fastq 2017 Mar 16 08:49:15 2.01 GB
Heat2_2.fastq 2017 Mar 16 08:37:16 2.01 GB
Heat3_1.fastq 2017 Mar 16 08:22:46 2.92 GB
Heat3_2.fastq 2017 Mar 16 09:03:59 2.92 GB
class_fastq
SRA_fastq
scripts
Community Data
Shared With Me
Trash
Favorites
Displaying 1 - 12 of 12
3 item(s)
Selected file:
Heat1_1.fastq, Heat2_1.fastq, Heat3_1.fastq
OK
Cancel

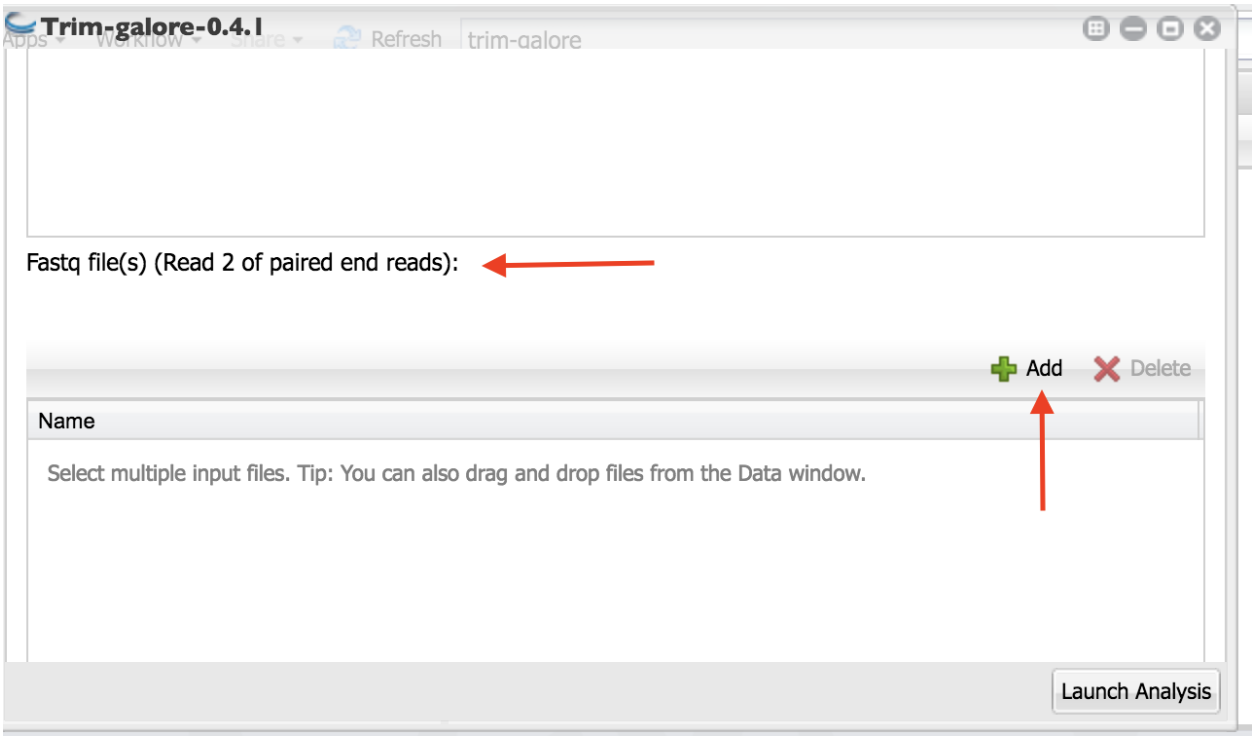


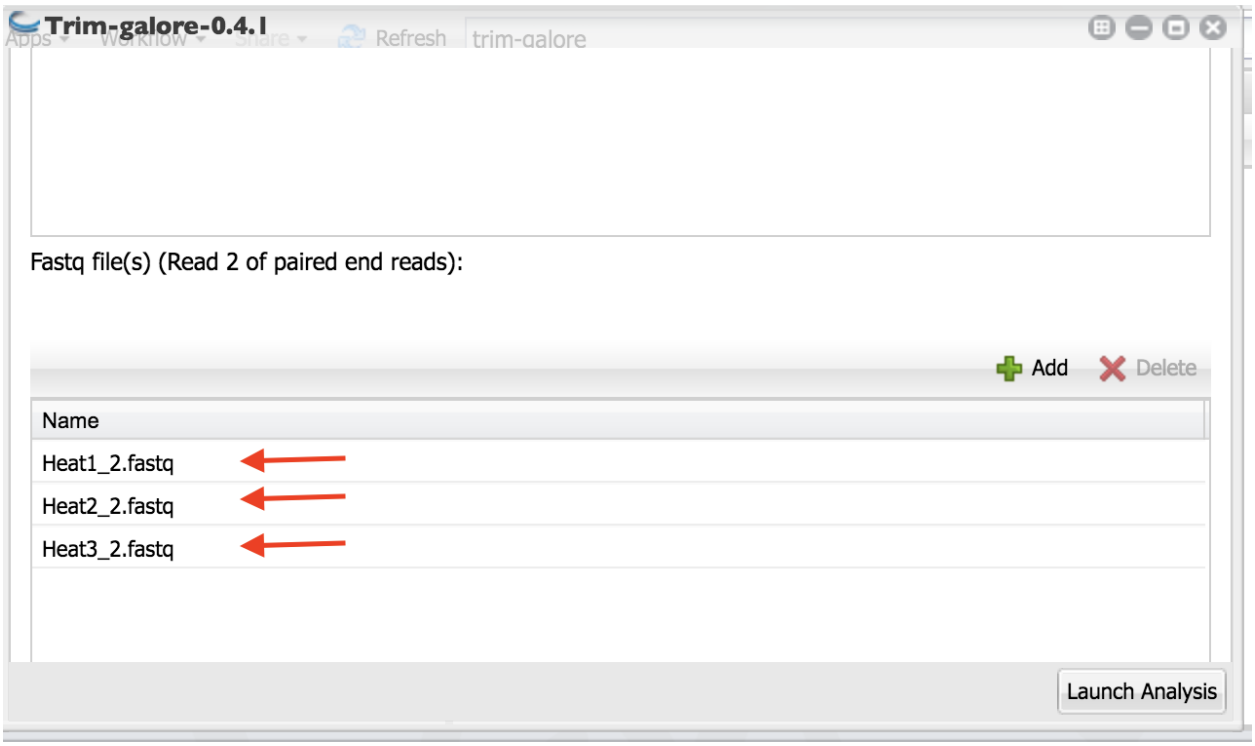
4. Scroll down and click on the "+" below "Fastq file(s) (Read 2 of paired end reads):"

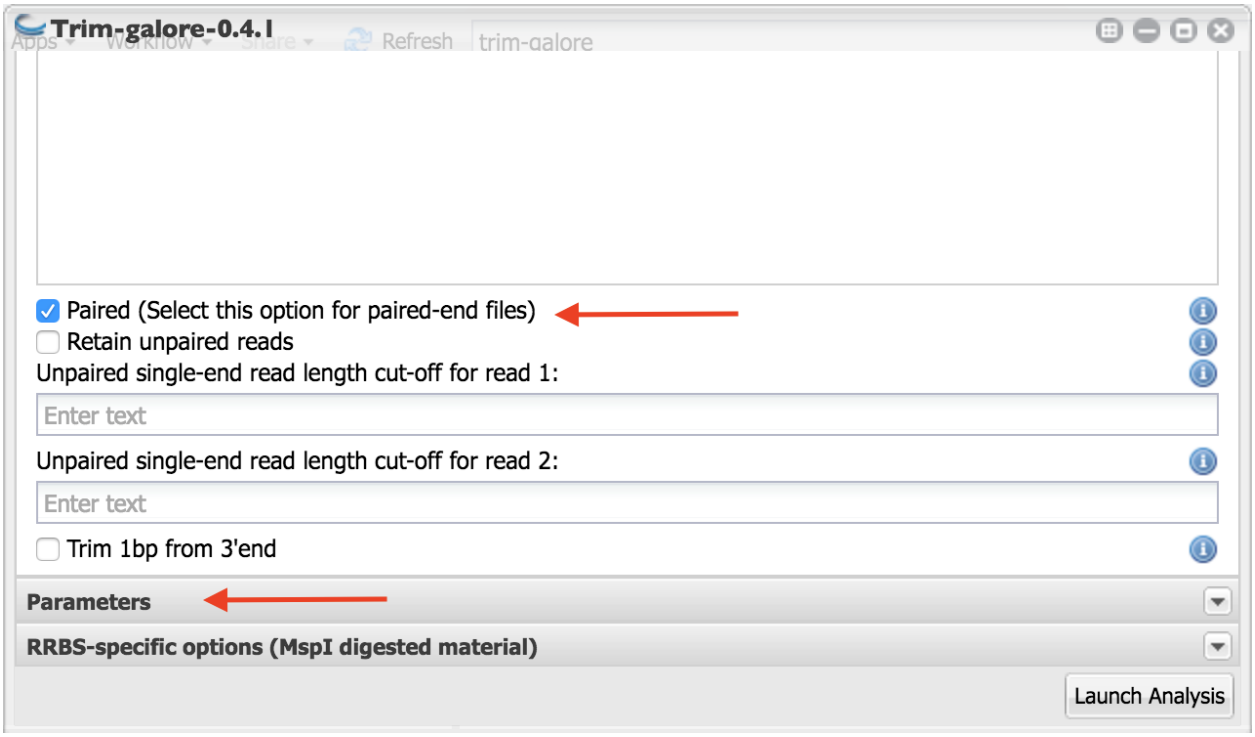
5. Select the read two files as above. You will see them in the box as in the figure below.

6. Scroll down and check box beside "Paired (Select this option for paired-end files)" to indicate these are paired end reads.

very important


Trim-galore-0.4.1
Refresh
trim-galore
Fastq file(s) (Read 2 of paired end reads):
Add
Delete
Name
Select multiple input files. Tip: You can also drag and drop files from the Data window.
Launch Analysis


Trim-galore-0.4.1
Refresh
trim-galore
Fastq file(s) (Read 2 of paired end reads):
Add
Delete
Name
Heat1_2.fastq
Heat2_2.fastq
Heat3_2.fastq
Launch Analysis



7. Click on "Parameters" as indicated in the above figure.

8. Set the parameters as indicated in the figure:

a. Use Fred 20 as quality trimming cut off (this is the default).

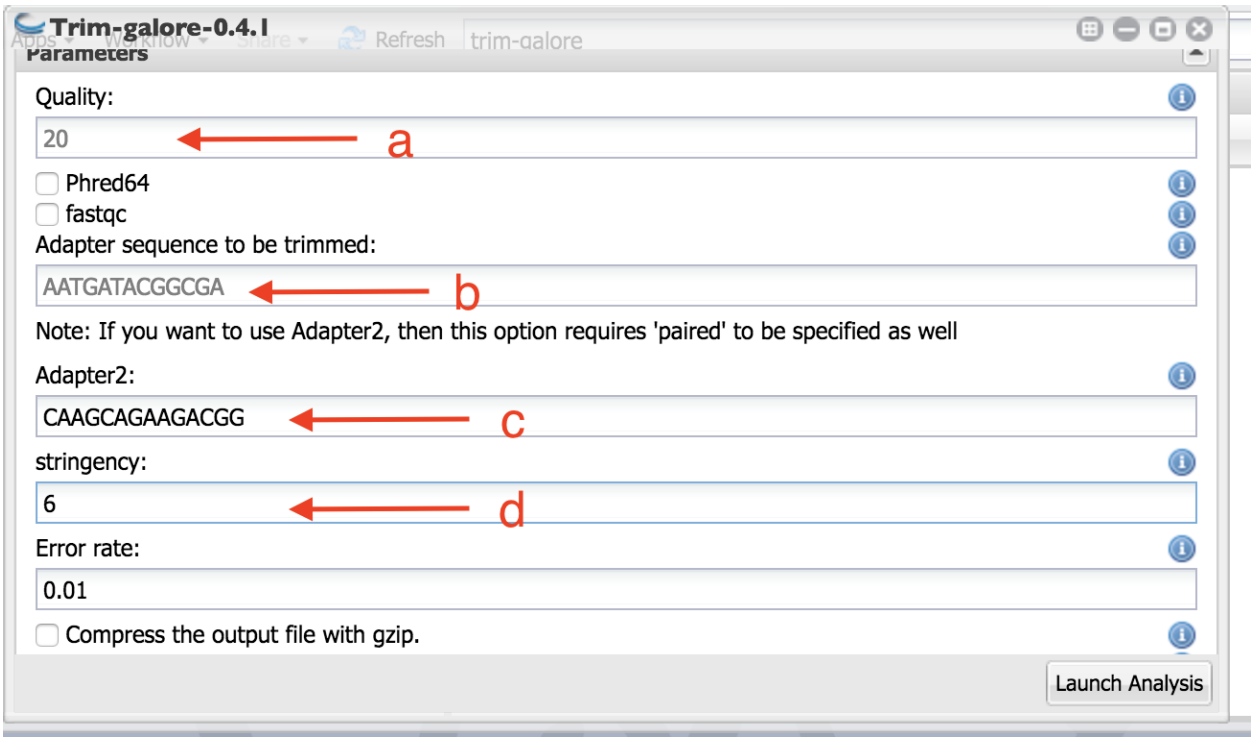
b. Copy and paste the following adapter sequence for in the box below "Adapter sequence to be trimmed:"

AATGATACGGCGA

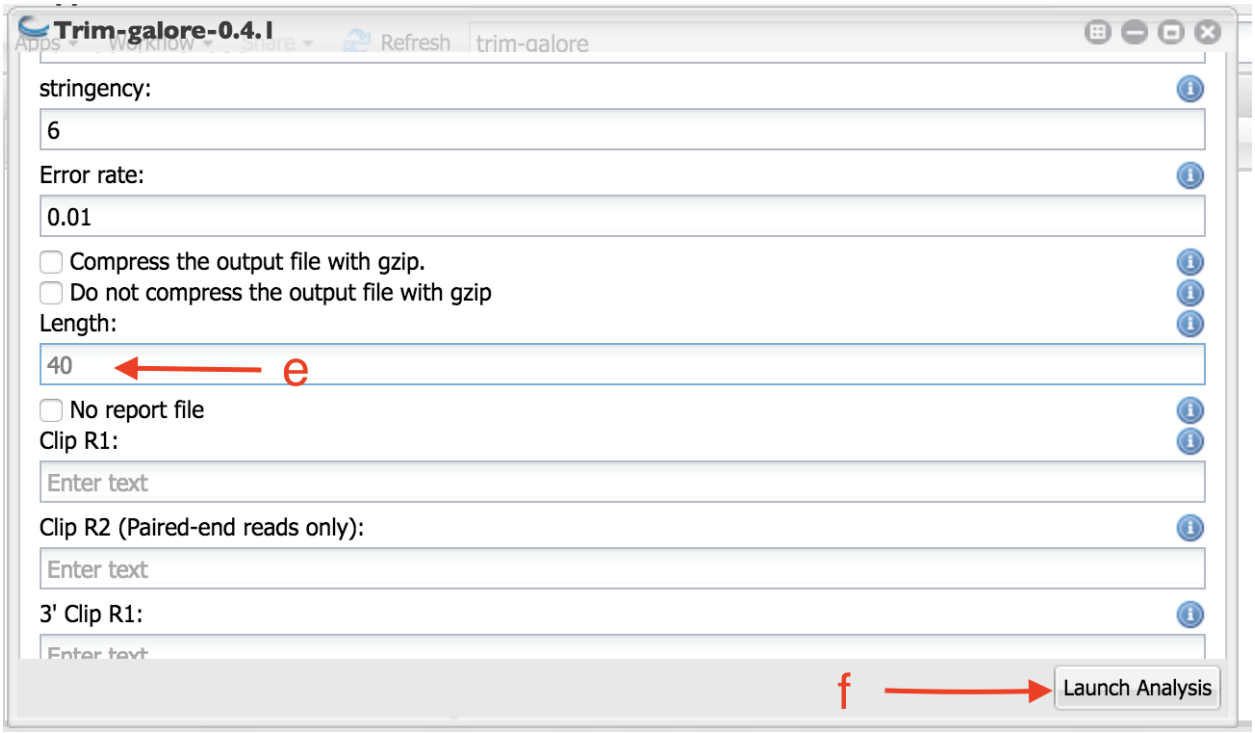
c. Copy and paste the following adapter sequence for in the box below "Adapter2"

CAAGCAGAAGACGG

d. Set the stringency to 6.



e. Scroll down and set the length as 40. Any sequence become shorter than this length during the trimming will be discarded. | f. Launch the analysis.


Trim-galore-0.4.1
Refresh
trim-galore
stringency:
6
Error rate:
0.01
Compress the output file with gzip.
Do not compress the output file with gzip
Length:
40
e
No report file
Clip R1:
Enter text
Clip R2 (Paired-end reads only):
Enter text
3' Clip R1:
f
Launch Analysis

CHAPTER 7

Mapping short reads

If you are using genome as the reference for RNAseq reads, you will need to use a splice-aware aligner like Tophat2. If you are using cDNA as the reference, you can use a general purpose aligner like Bowtie2.

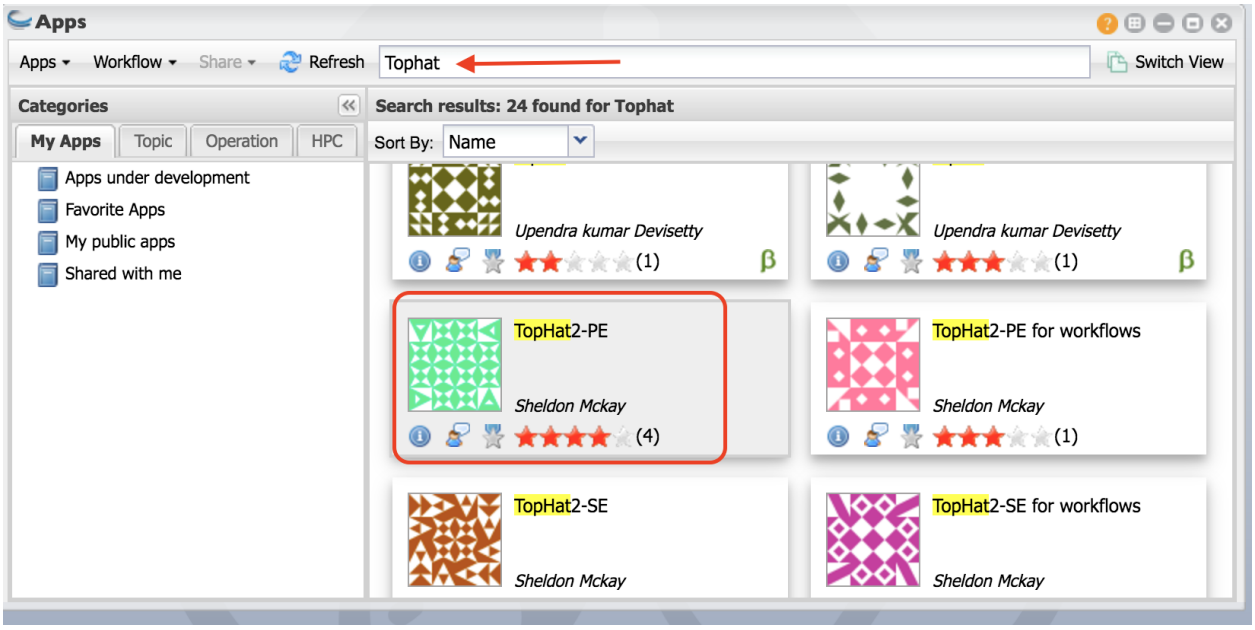
You need to do only one of the procedures based on what your group have been assigned to.

7.1 Step 1: Mapping with Tophat2

1. Click on *App*.

2. In the finder window type "Tophat"

3. Select "Tophat2-PE".



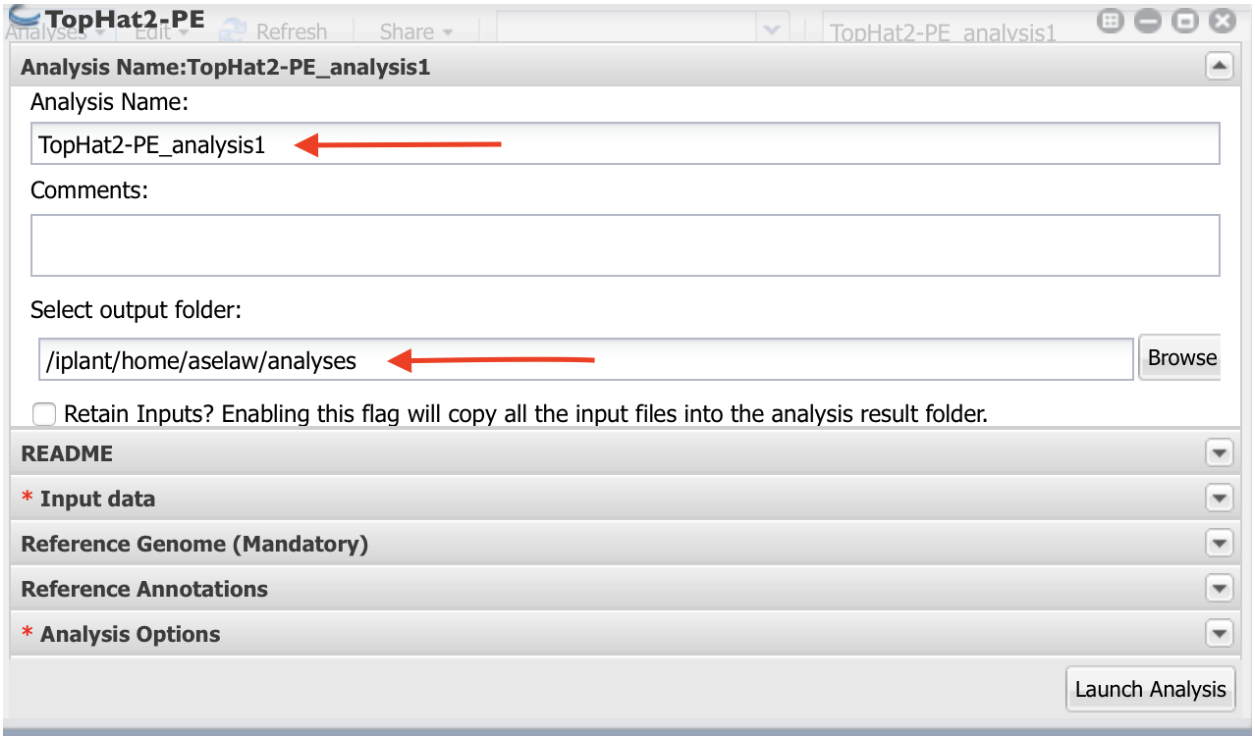
4. As indicated in the figure, Name your analysis as you want.

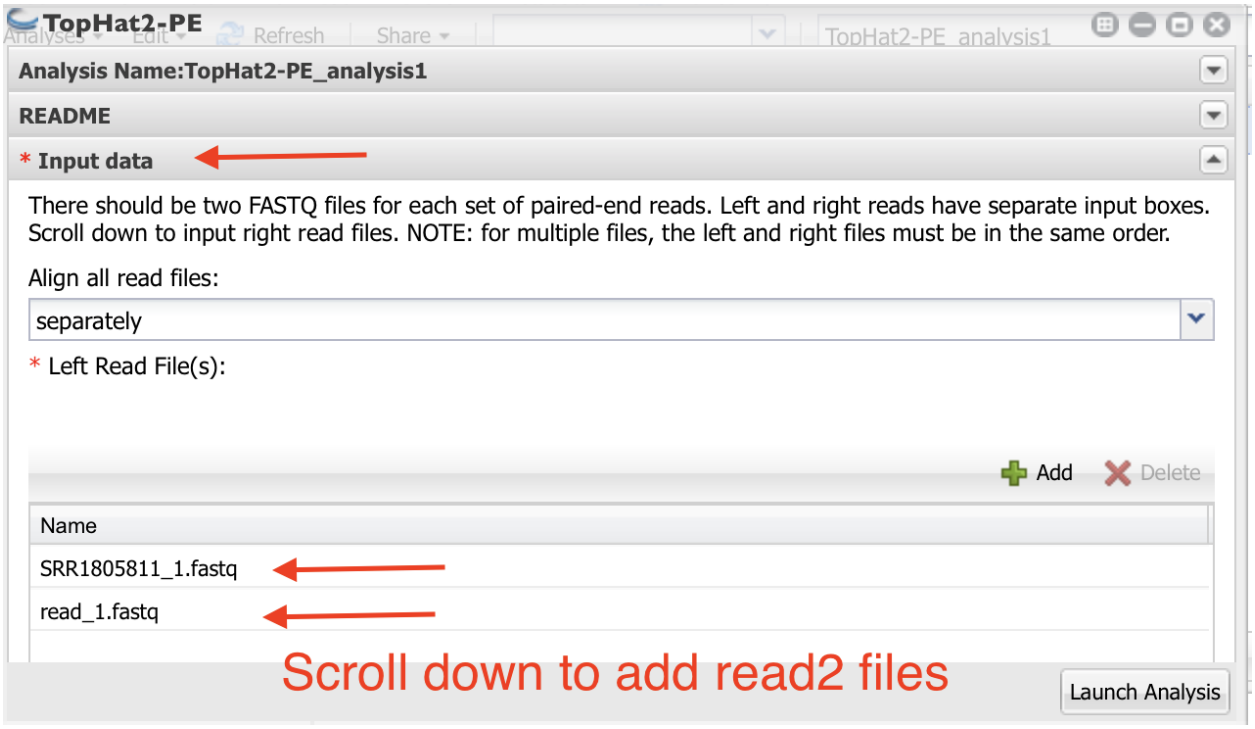
5. Select the output folder where your analysis is going to be.

6. Click on “Input data”

7. Click on the Green “+” sign.

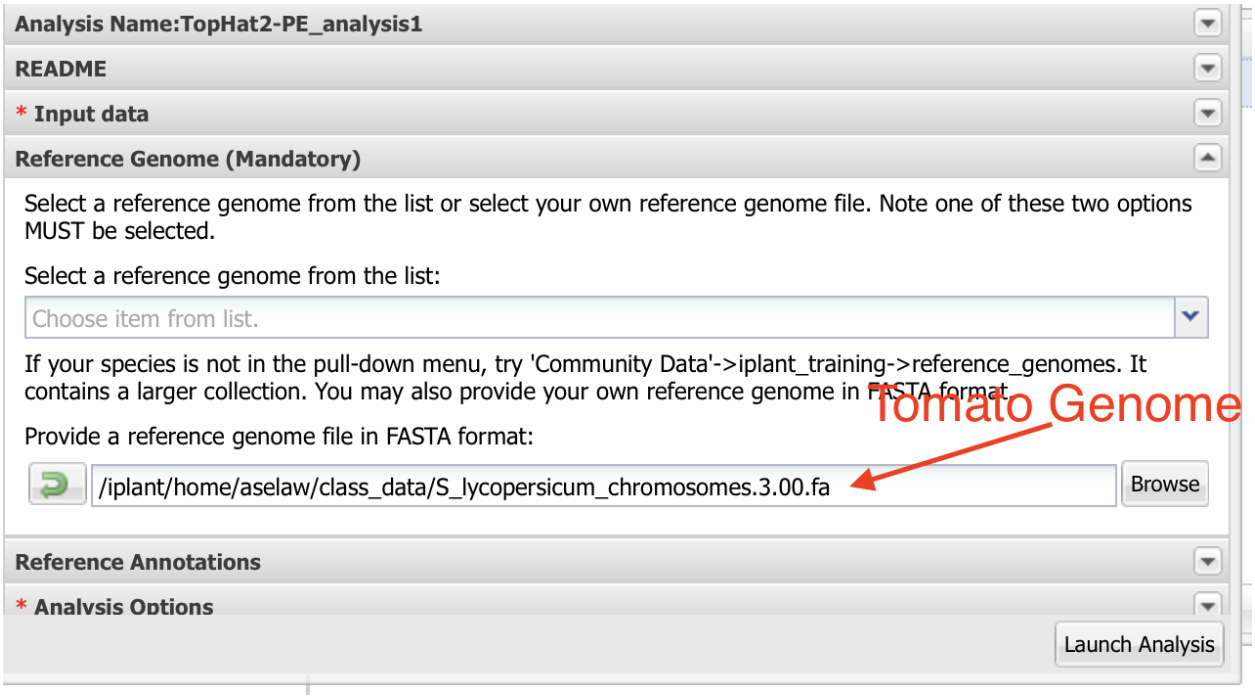
8. Navigate to the folder where your samples are located. Select only the **first read files**. Click “OK”. **You can select all three of your first read files.**


TopHat2-PE
Refresh
Share
TopHat2-PE_analysis1
Analysis Name:TopHat2-PE_analysis1
Analysis Name:
TopHat2-PE_analysis1
Comments:
Select output folder:
/iplant/home/aselaw/analyses
Browse
Retain Inputs? Enabling this flag will copy all the input files into the analysis result folder.
README
* Input data
Reference Genome (Mandatory)
Reference Annotations
* Analysis Options
Launch Analysis

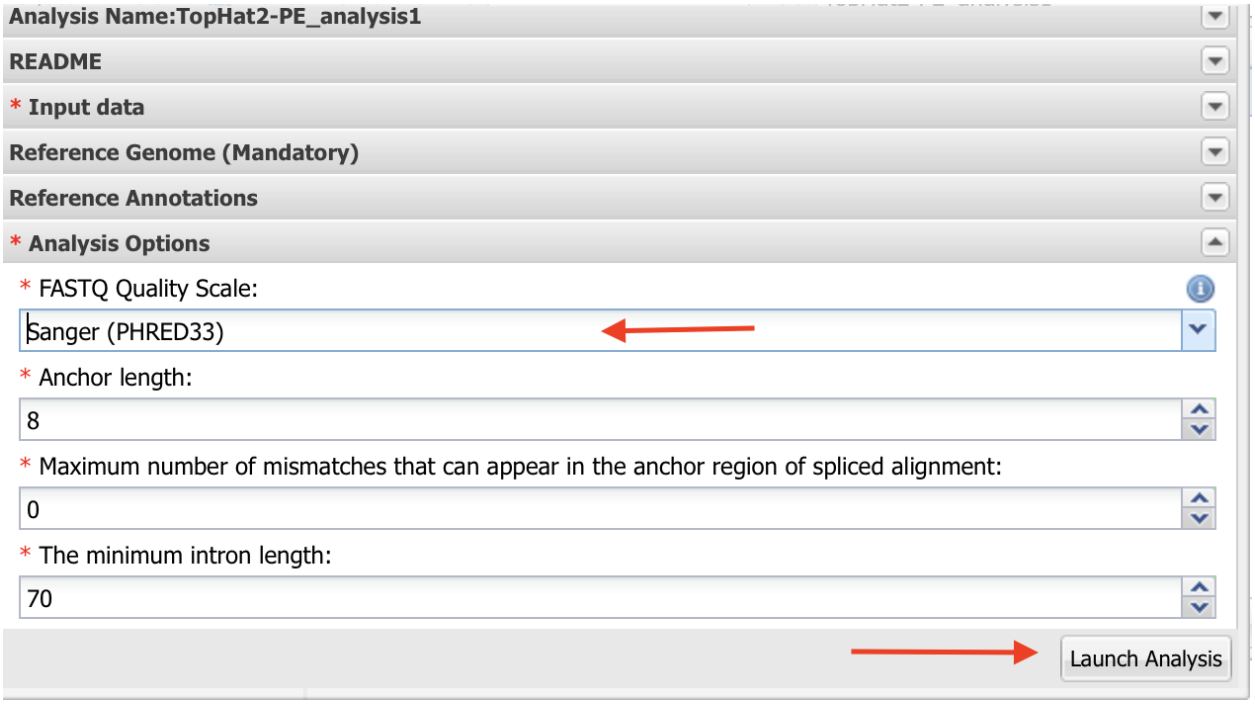

TopHat2-PE
Refresh
Share
TopHat2-PE_analysis1
Analysis Name:TopHat2-PE_analysis1
README
* Input data
There should be two FASTQ files for each set of paired-end reads. Left and right reads have separate input boxes.
Scroll down to input right read files. NOTE: for multiple files, the left and right files must be in the same order.
Align all read files:
separately
* Left Read File(s):
Add
Delete
Name
SRR1805811_1.fastq
read_1.fastq
Scroll down to add read2 files
Launch Analysis

9. Scroll down and click on the "+" below "Fastq file(s) (Read 2 of paired end reads):"

10. Select "Reference Genome" and select the tomato genome sequence as input.



11. Make sure quality is Sanger and leave rest of the default values as they are. Launch the analysis.



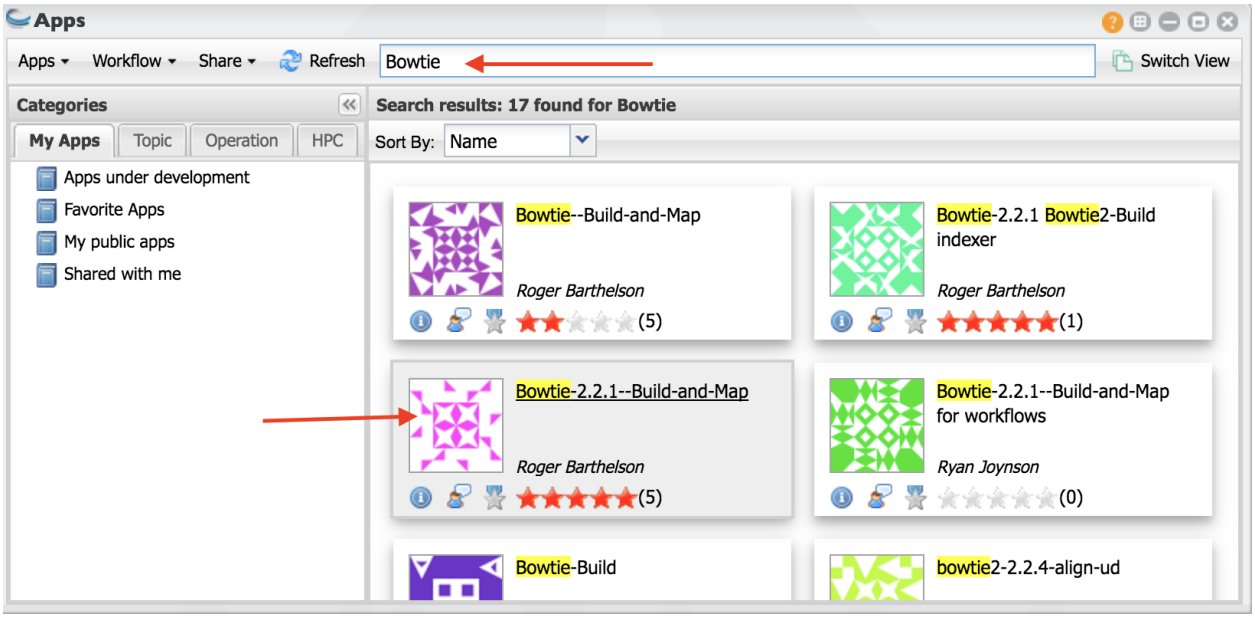
7.2 Step 2: Mapping with Bowtie2

1. Click on *App*.

2. In the finder window type "Bowtie".

3. Select Bowtie app indicated in the figure.

4. As indicated in the figure, Name your analysis as you want.

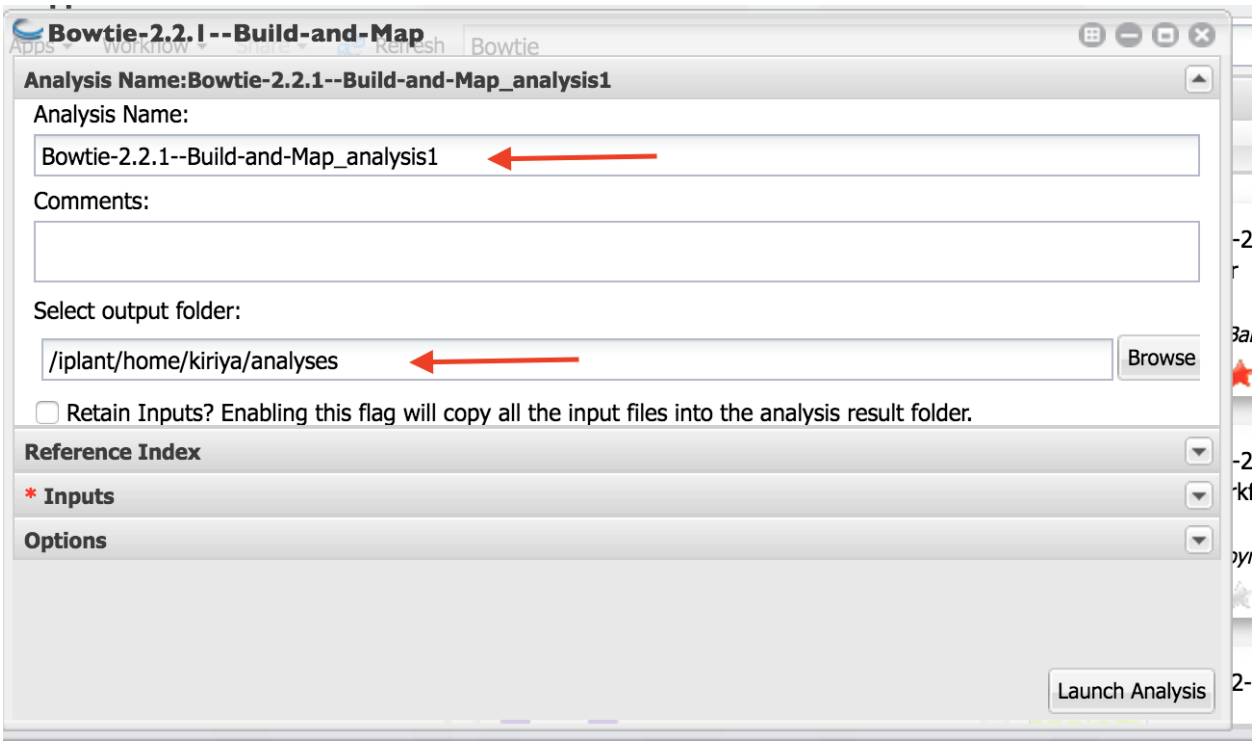


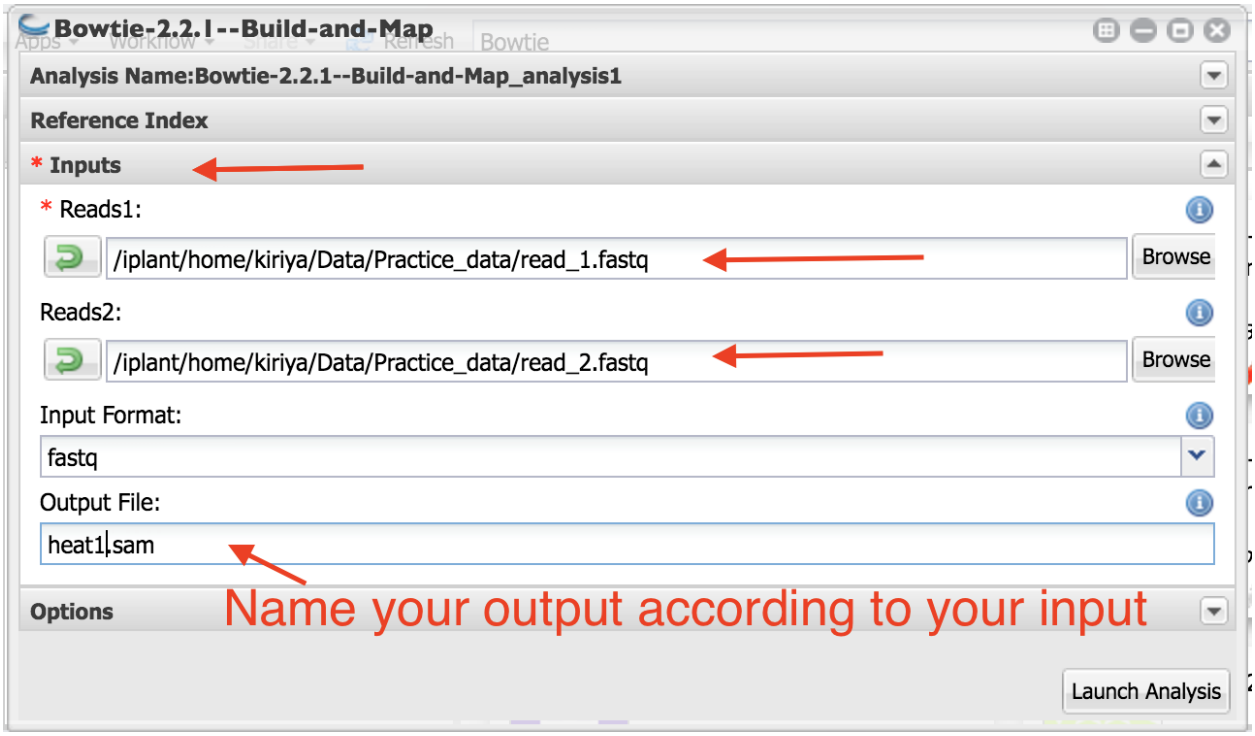
5. Select the output folder where your analysis is going to be.

6. Click on “Input”

8. Navigate to the folder where your samples are located. Select first and second read files. You can only input one sample at a time.

9. You need to name your output file carefully. For e.g., if it is heat1 sample, name the output as heat1.sam.

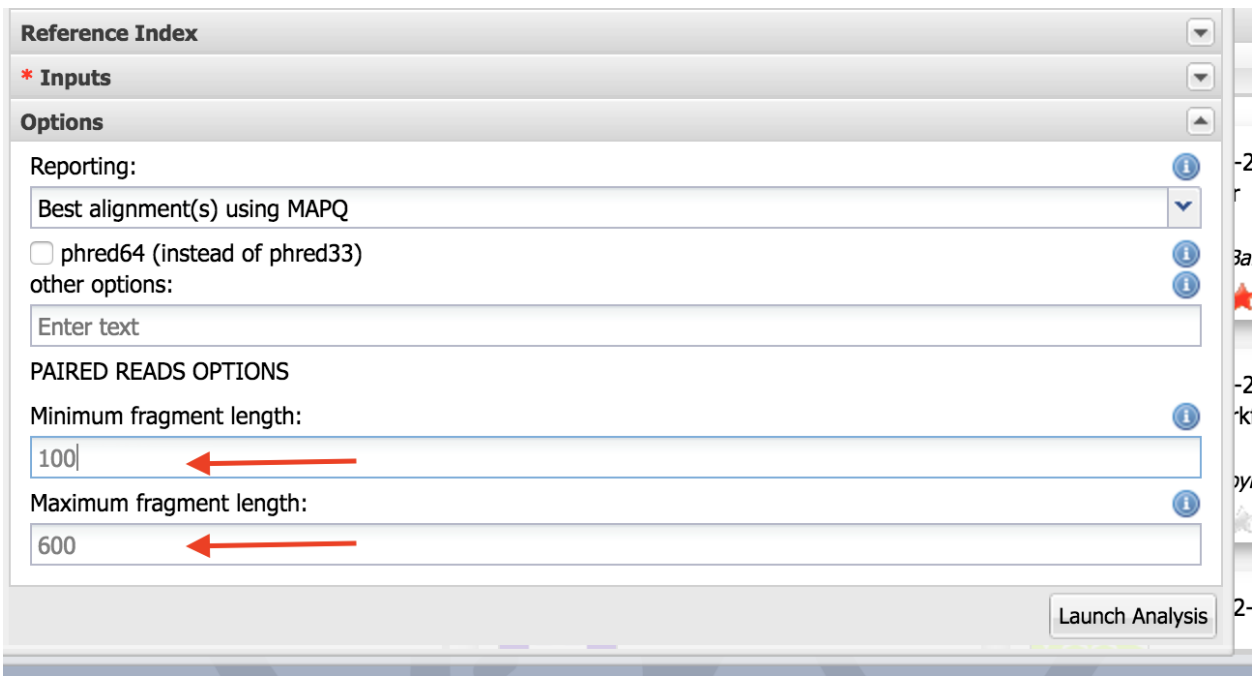

Bowtie-2.2.1--Build-and-Map
Analysis Name:Bowtie-2.2.1--Build-and-Map_analysis1
Analysis Name:
Bowtie-2.2.1--Build-and-Map_analysis1
Comments:
Select output folder:
/iplant/home/kiriya/analyses
Browse
Retain Inputs? Enabling this flag will copy all the input files into the analysis result folder.
Reference Index
* Inputs
Options
Launch Analysis


Bowtie-2.2.1--Build-and-Map
Analysis Name:Bowtie-2.2.1--Build-and-Map_analysis1
Reference Index
* Inputs
* Reads1:
/iplant/home/kiriya/Data/Practice_data/read_1.fastq
Browse
Reads2:
/iplant/home/kiriya/Data/Practice_data/read_2.fastq
Browse
Input Format:
fastq
Output File:
heat1.sam
Name your output according to your input
Options
Launch Analysis

10. Select “Reference Index” and select the tomato cDNA sequence as input.

img/bowtie_3.png

11. Select options. Set “Minimum fragment length” as 100 and “Maximum fragment length” as 600. Launch the analysis.



CHAPTER 8

Counting mapped reads

To get the number of reads mapped to a reference sequences (in this case, predicted tomato cDNA sequences), we can use Samtools. Bowtie2 output is in sam format and first, we need to convert the output files into sorted bam files.

1. Type Samtools in app finding window.
2. Select "SAM to sorted BAM"
3. Select Bowtie2 output files (SAM format).

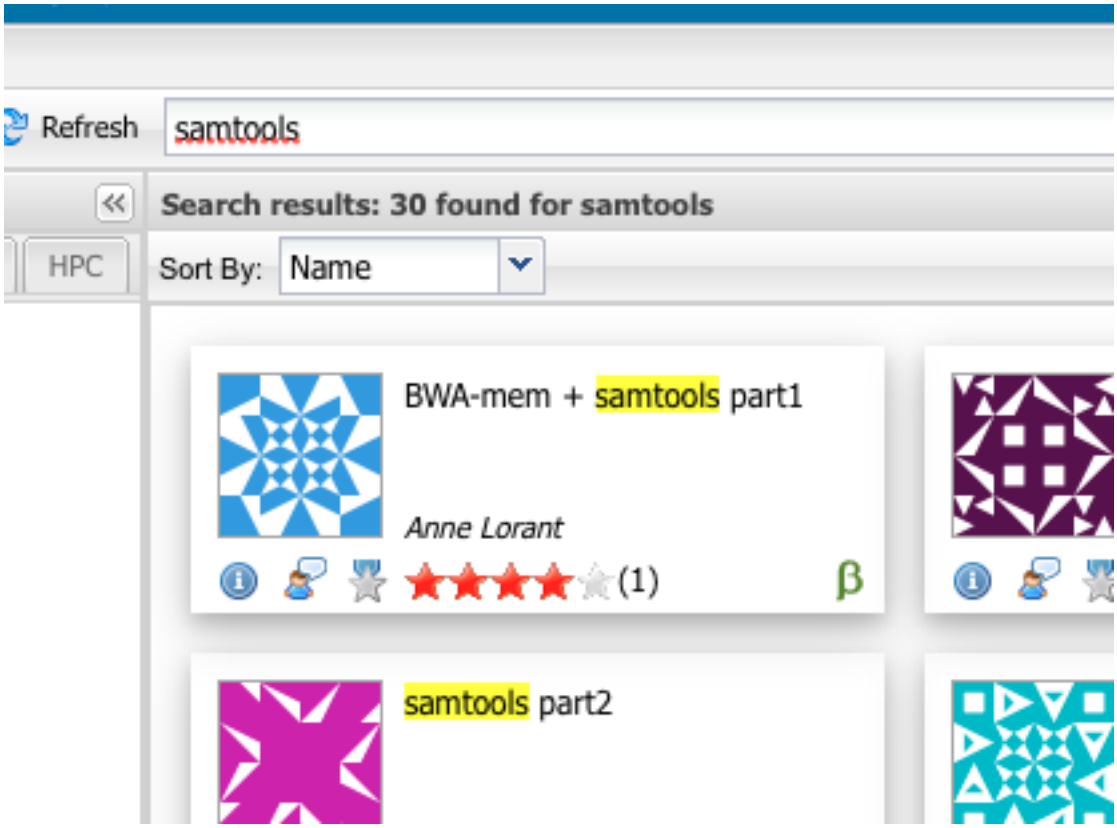
4. Above will create sorted bam file. You will need to use this as the input for the Samtools Flagstat, which will count the number of mapped reads.

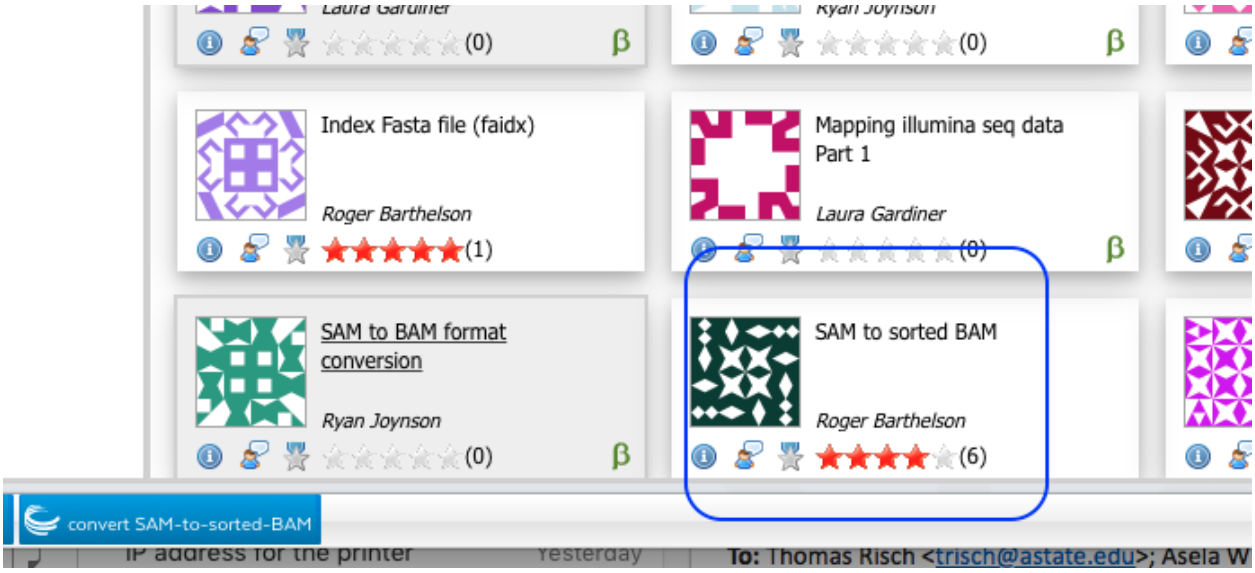
** You can get the flagstat for all six files from following link.** | https://github.com/ajwije/2017_spring_Bioinfo_class/blob/master/Files/flagstat.txt

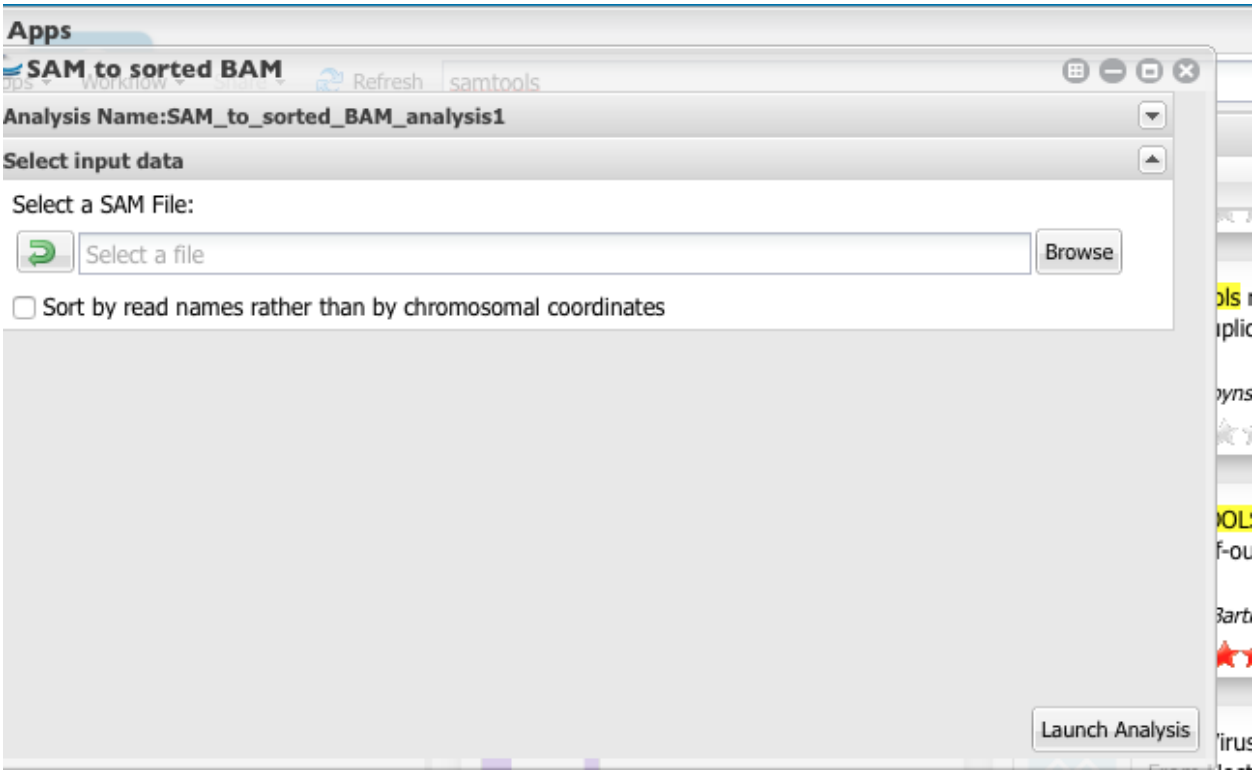
I have used the following bash command to count mapped reads in case you are interested in it doing programmatically.

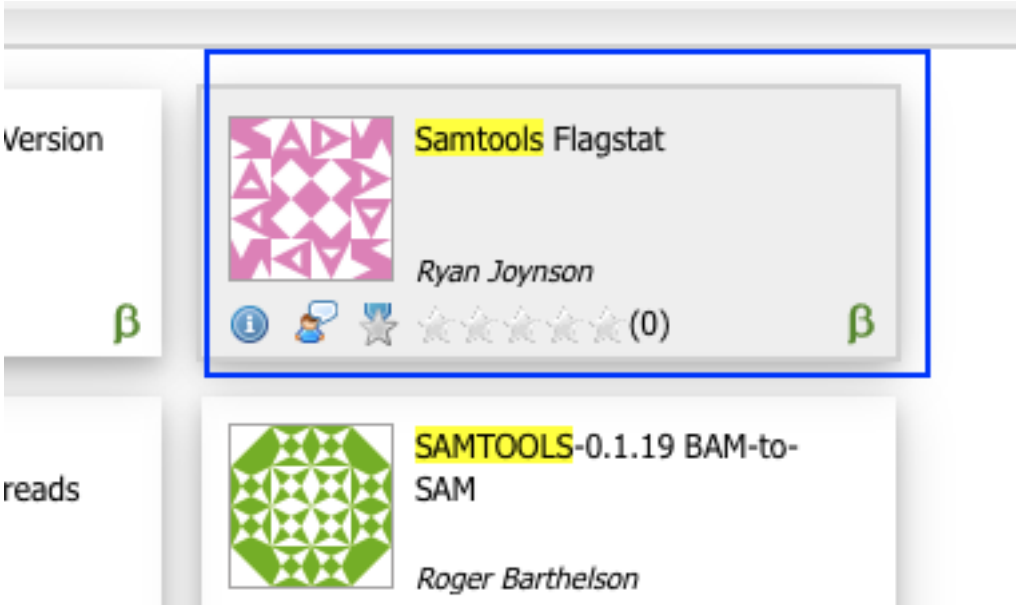
```
# navigate to where your sam files are and execute the following␣
↪commands.

for i in *.sam
```

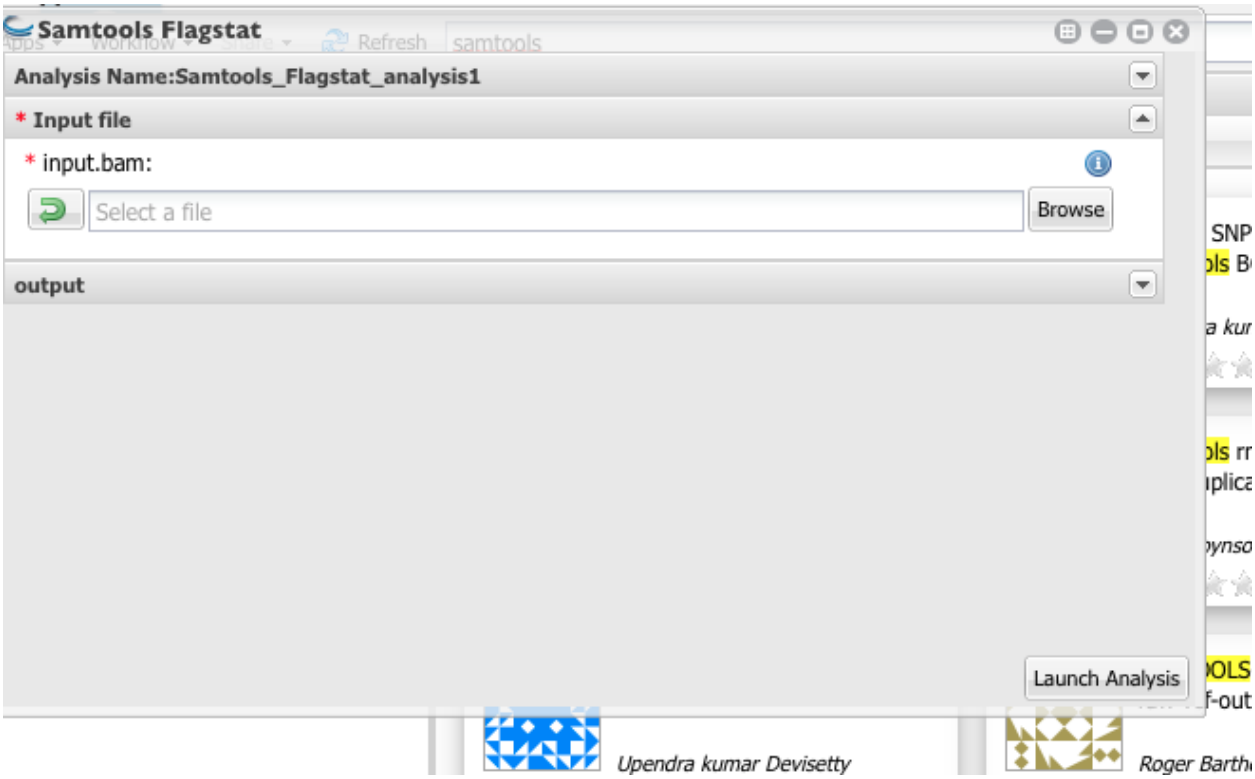
(continues on next page)


Refresh
samtools
Search results: 30 found for samtools
HPC
Sort By:
Name
BWA-mem + samtools part1
Anne Lorant
(1)
samtools part2


(0)
(0)
Index Fasta file (faidx)
Roger Barthelson
(1)
Mapping illumina seq data
Part 1
Laura Gardiner
(0)
SAM to BAM format
conversion
Ryan Joynson
(0)
SAM to sorted BAM
Roger Barthelson
(6)
convert SAM-to-sorted-BAM
To: Thomas Risch <trisch@astate.edu>; Asela W


Apps
SAM to sorted BAM
Refresh
samtools
Analysis Name:SAM_to_sorted_BAM_analysis1
Select input data
Select a SAM File:
Select a file
Browse
Sort by read names rather than by chromosomal coordinates
Launch Analysis


Version
Samtools Flagstat
Ryan Joynson
(0)
reads
SAMTOOLS-0.1.19 BAM-to-SAM
Roger Barthelson



(continued from previous page)

```
do
#extract the file name without extension and print it to the screen


echo ${i%.sam}

#covert sam to sorted bam
samtools view -bS $i | samtools sort - -o ${i%.sam}_sorted.bam

#getting flagstat and write it an output file for each bam file.
samtools flagstat ${fname%.sam}_sorted.bam >> flagstat.txt

done
```

CHAPTER 9

Differential gene expression analysis

Link for Bowtie mapped counts http://de.cyverse.org/dl/d/E9B4C299-D6CB-4656-A4F6-FF67240AEA49/170407_bowtie_counts.txt

Target file for bowtie mapped reads: http://de.cyverse.org/dl/d/BECB62C3-A369-4084-9BC9-2BFD9E6E9600/bowtie_target.txt

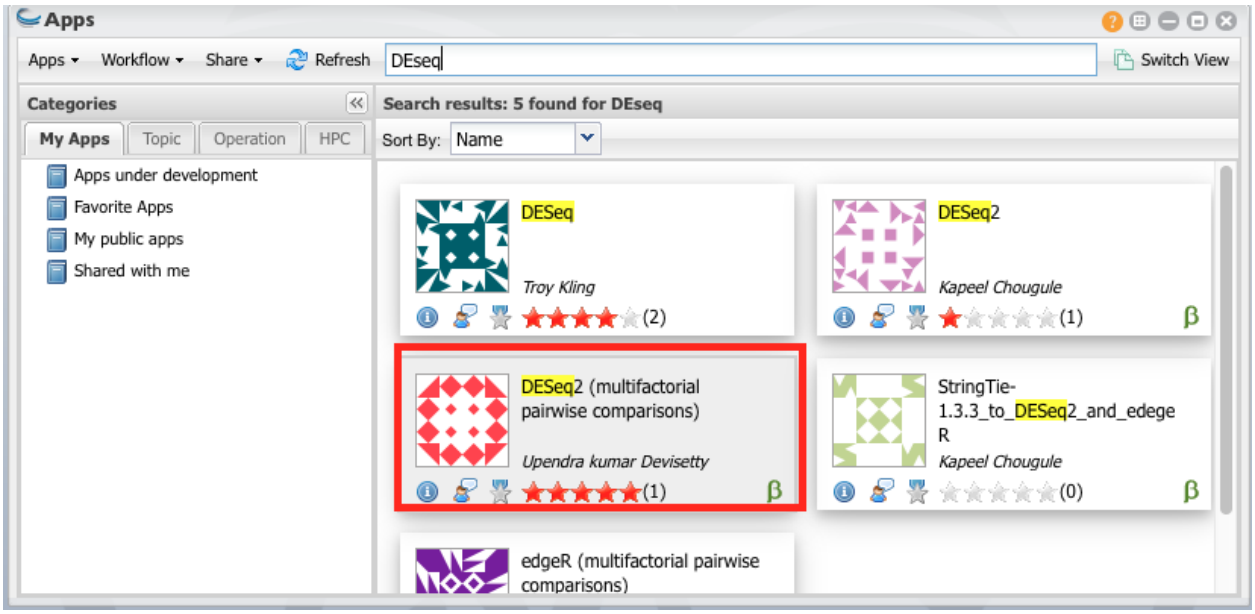
9.1 DESeq tutorial:

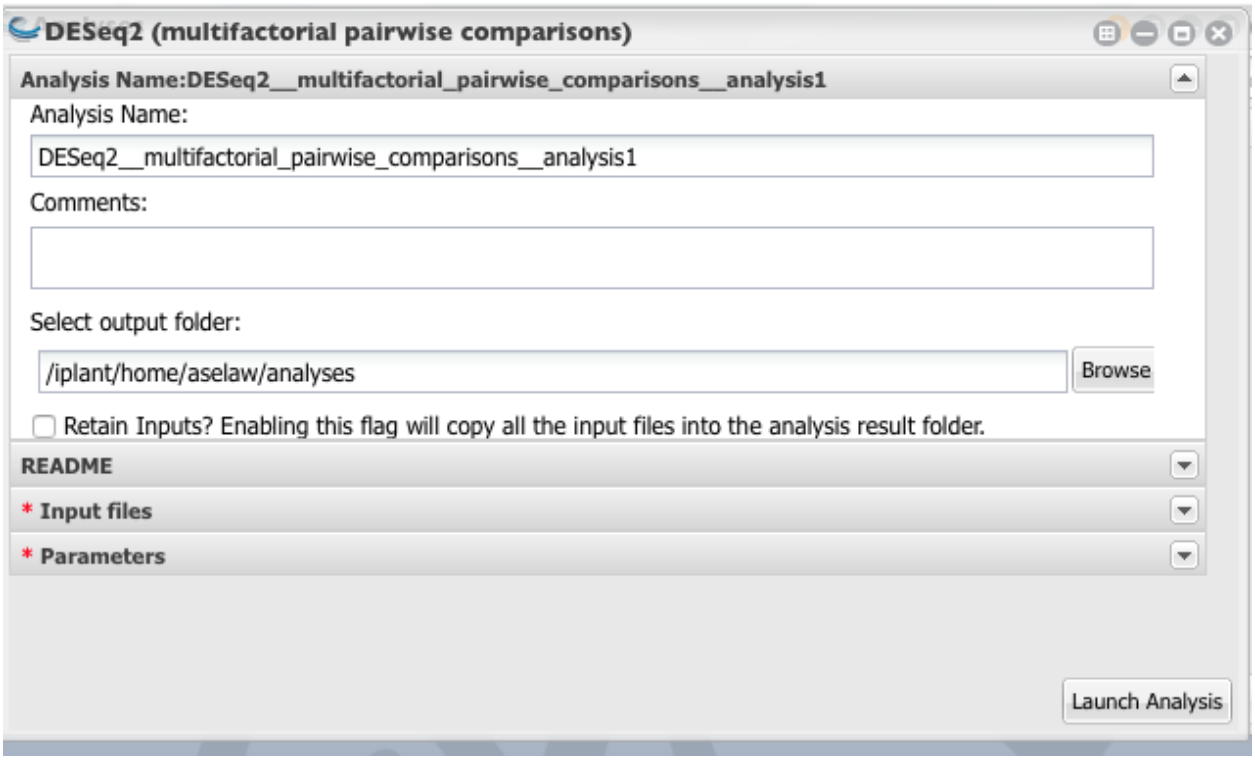
Tutorial link

9.2 Steps to perform DEseq analysis

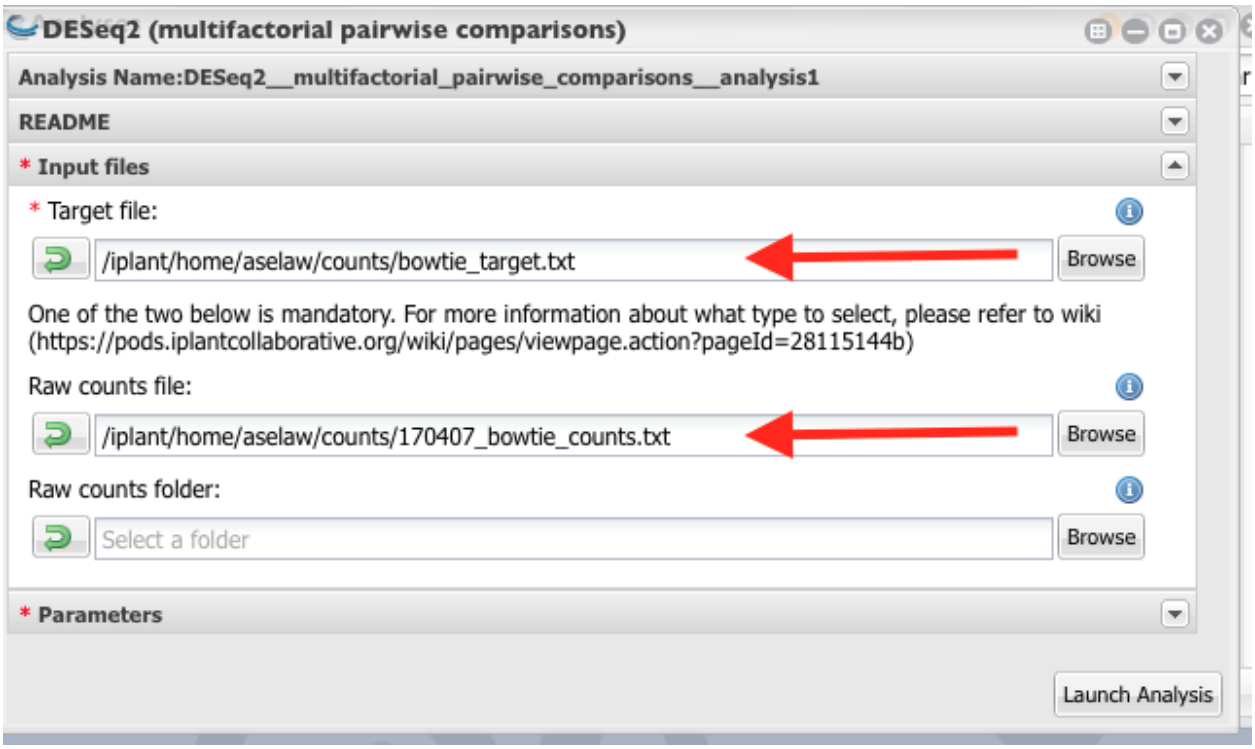
1. From Apps select "DEseq (Multifactorial Comparison)

2. Name your analysis and select a folder where your results need to be saved.


Apps
Apps
Workflow
Share
Refresh
DEseq
Switch View
Categories
Search results: 5 found for DEseq
My Apps
Topic
Operation
HPC
Sort By:
Name
Apps under development
Favorite Apps
My public apps
Shared with me
DESeq
Troy Kling
(2)
DESeq2
Kapeel Chougule
(1)
DESeq2 (multifactorial pairwise comparisons)
Upendra kumar Devisetty
(1)
StringTie-1.3.3_to_DESeq2_and_edegeR
Kapeel Chougule
(0)
edgeR (multifactorial pairwise comparisons)

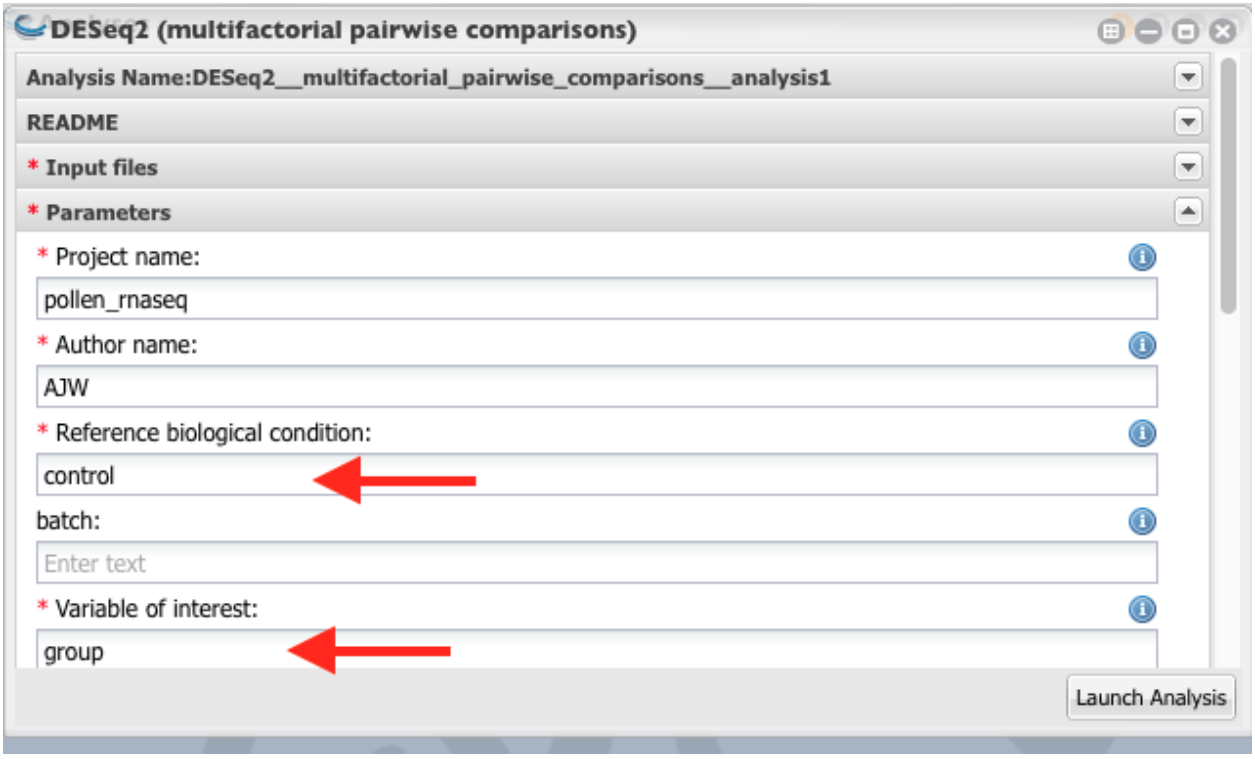

DESeq2 (multifactorial pairwise comparisons)
Analysis Name:DESeq2__multifactorial_pairwise_comparisons__analysis1
Analysis Name:
DESeq2__multifactorial_pairwise_comparisons__analysis1
Comments:
Select output folder:
/iplant/home/aselaw/analyses
Browse
Retain Inputs? Enabling this flag will copy all the input files into the analysis result folder.
README
* Input files
* Parameters
Launch Analysis

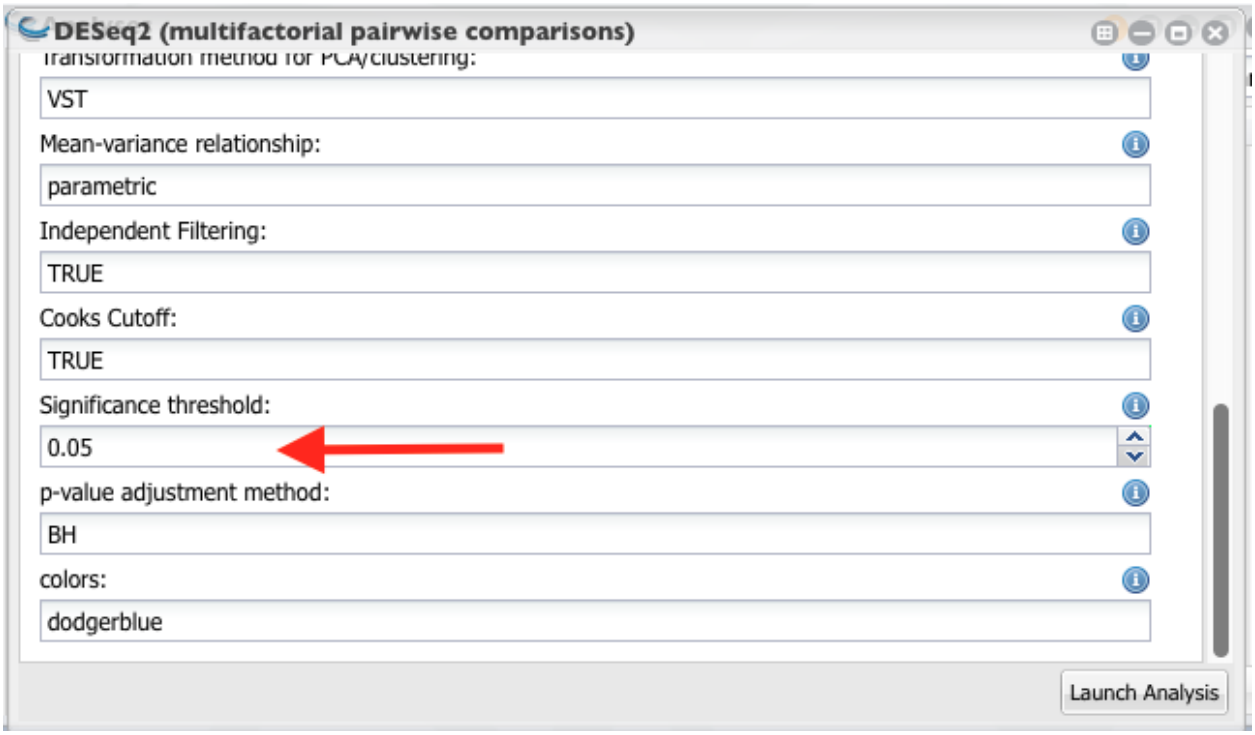
3. Select the correct target file and the count file.



4. Give a name to the project. Reference biological condition should be "control" samples. Variable of interest is "group" (Column header of the third column of the target file).

5. Set the significant threshold to 0.05 and launch the analysis.


DESeq2 (multifactorial pairwise comparisons)
Analysis Name:DESeq2__multifactorial_pairwise_comparisons__analysis1
README
* Input files
* Parameters
* Project name:
pollen_rnaseq
* Author name:
AJW
* Reference biological condition:
control
batch:
Enter text
* Variable of interest:
group
Launch Analysis


DESeq2 (multifactorial pairwise comparisons)
VST
Mean-variance relationship:
parametric
Independent Filtering:
TRUE
Cooks Cutoff:
TRUE
Significance threshold:
0.05
p-value adjustment method:
BH
colors:
dodgerblue
Launch Analysis

9.3 DE gene list

I have used the following R code to merge the DE genes list and the functions.

```
library(reshape2)
library(readr)

# Used the terminal command to grep the fasta headers and wrote it to␣
↪a file called "ITAG3_10_cDN_names.txt"
#imported this file to Rstudio

# Removed the ">" sign
ITAG3_10_formated_names <- as.data.frame(sapply(ITAG3_10_cDN_names,␣
↪gsub, pattern = ">", replacement = "" ))

#Seperate gene ids and description using space as delimiter
ITAG3_10_formated_names <- data.frame(colsplit(ITAG3_10_formated_names
↪$X1, " ", c("Id", "Description")))

#imported up regulated genes to Rstudio and merge with the above file␣
↪using gene ids.
heatvscontrol_up_func <- merge(heatvscontrol_up, ITAG3_10_formated_
↪names,
                       by.x = "Id",
                       by.y = "Id")

#write output
write.table(x = heatvscontrol_up_func, file = "heatvscontrol_up_func.
↪txt", quote = FALSE, sep = "\t", row.names = FALSE)


#imported down regulated genes to Rstudio and merge with the above␣
↪file using gene ids
heatvscontrol_down_func <- merge(heatvscontrol_down, ITAG3_10_formated_
↪names,
                       by.x = "Id",
                       by.y = "Id")
#write output
write.table(x = heatvscontrol_down_func, file = "heatvscontrol_down_
↪func.txt", quote = FALSE, sep = "\t", row.names = FALSE)
```

Up-regulated gene list: http://de.cyverse.org/dl/d/E641698E-8688-4C20-B829-0B12BABC8ABB/heatvscontrol_up_func.txt

Down-regulated gene list: http://de.cyverse.org/dl/d/3C45B913-612F-4B97-8F44-8021470AE527/heatvscontrol_down_func.txt

CHAPTER 10

Secondary Structure Prediction

1. We will use one of the differentially expressed in tomato pollen transcriptome under head stress.

I have retrive the amino acid sequences for Solyc06g050510 from SolGen website.

MKRHIHYNAHPIDPHPFEAFWYGSWQAVERLRINMGTITTHVLVDGEVIEENIPVTNLRMRSRKATLSDC
FLRPGLEVCVLSIPYQGENSGDEKDVKPVWIDGKIRSIERKPHELTCTCKFHVSVYVTQGPPPILKKTLSK
IKMLPIDQIAVLQKLEPKPCENKRYRWSSSEDCNSLQTFKLFIGKFSSDLTWLM-
TASVLKEATFDVRSIHNQ IVYEIVDDDLVRKETNSNQHSYSVNFKLEGGVQTTTVIQFN-
RDIPDINSTSDLSESGPLVLYDLMGPRRSKR RFVQPERYYGCDDDMAEFDVEMTRLVG-
GRRKVEYEELPLALSIQADHAYRTGEIEEISSSYKRELFGGNIRS HEKRSSESSSGWR-
NALKSDVNKLADKKSVTADRQHQLAIVPLHPPSGTGLTVHEQVPLDVDVPEHLSAEIGE
IVSRYIHFNSSSTSHDRKASKMNFTKPEARRWGQVKISKLKFMGLDRRGGTL-
GSHKKYKRNTTKKDSIYDIR SFKKGSVAANVYKELIRRCMANIDATLNKEQPPI-
IDQWKEFQSTKSSQRESGDHLAMNRDEEVSEIDMLWKE MELALASCYLLDDSED-
SHAQYASNVRIGAEIRGEVCRHDYRLNEEIGIICRLCGFVSTEIKDVPPPFMPSSN
HNSSKEQRTEEATDHKQDDDGLDTLSIPVSSRAPSSSGGGEGNVWALIPDL-
GNKLRVHQKRAFEFLWKNIAG SIVPAEMQPESKERGGCVISHTPGAGKTL-
LIISFLVSYLKLFPGSRPLVLAPKTTLYTWYKEVLKWKIPVPV YQIHGGQT-
FKGEVLREKVKLCPGLPRNQDVMHVLDCLEKMQMWLSQPSVLLMGYTSFLTL-
TREDSPYAHRKY MAQVLRQCGLLILDEGHNPRSTKSRLRKGLMKVNTRLRILLS-
GTLFQNNFGEYFNTLTLARPTFVDEVLKEL DPKYKNKNKGASRFSLENRARKM-

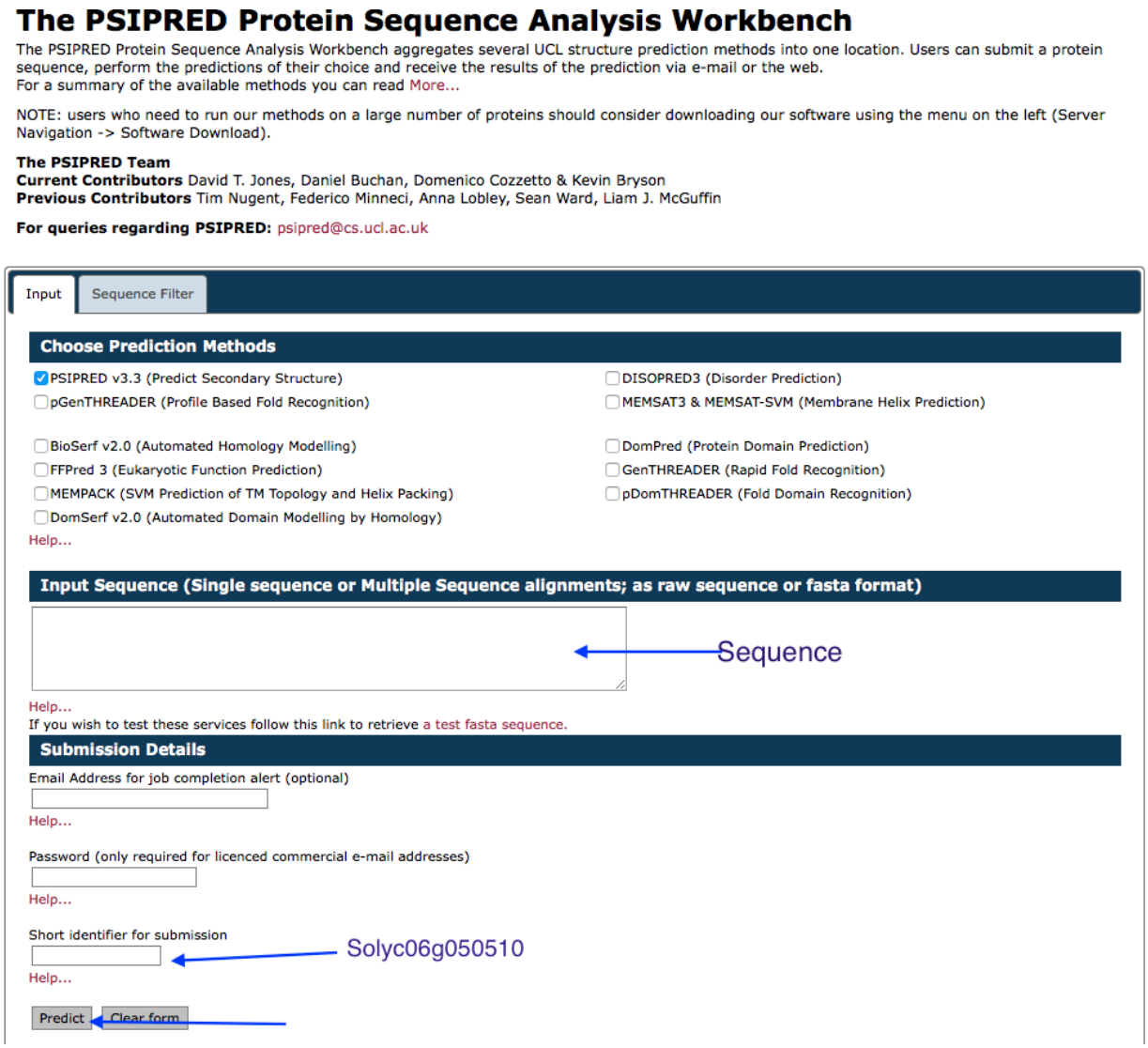
FIDKISTVIDSDIPKKRKEGLNILKKLTGGFIDVHDGGTSDNLPGLQCY TLMMK-
STTLQQEILVKLQNQRPIYKGFPLELELLITLGAIHPWLIRTTACSSQYFKEEE-
LEALQKFKFDLKL GSKVKFVMSLIPRCLLRREKVLIFCHNIAPINLFLEIFERFYG-
WRKGIEVLVLQGDIELFQRGRIMDLFEEP GGPSKVMLASITTCAEGISLTAASRVILLD-
SEWNPSKSKQAIARAFRPGQDKVVYVYQLLATGTLEEEKYKR TTWKEWVSSMIFS-
EDLVEDPSHWQAPKIEDELLREIVEEDRATLFHAIMKNEKASNMGSLQE

2. Point your browser to.

3. Copy and paste the amino acid sequence in the box and label the sequence.

4. We will use the previous submitted results:

http://bioinf.cs.ucl.ac.uk/psipred/result/e3f48c8e-28ff-11e7-879a-00163e110593


The PSIPRED Protein Sequence Analysis Workbench
The PSIPRED Protein Sequence Analysis Workbench aggregates several UCL structure prediction methods into one location. Users can submit a protein sequence, perform the predictions of their choice and receive the results of the prediction via e-mail or the web.
For a summary of the available methods you can read More...
NOTE: users who need to run our methods on a large number of proteins should consider downloading our software using the menu on the left (Server Navigation -> Software Download).
The PSIPRED Team
Current Contributors David T. Jones, Daniel Buchan, Domenico Cozzetto & Kevin Bryson
Previous Contributors Tim Nugent, Federico Minneci, Anna Lobley, Sean Ward, Liam J. McGuffin
For queries regarding PSIPRED: psipred@cs.ucl.ac.uk
Input
Sequence Filter
Choose Prediction Methods
PSIPRED v3.3 (Predict Secondary Structure)
pGenTHREADER (Profile Based Fold Recognition)
BioSerf v2.0 (Automated Homology Modelling)
FFPred 3 (Eukaryotic Function Prediction)
MEMPACK (SVM Prediction of TM Topology and Helix Packing)
DomSerf v2.0 (Automated Domain Modelling by Homology)
DISOPRED3 (Disorder Prediction)
MEMSAT3 & MEMSAT-SVM (Membrane Helix Prediction)
DomPred (Protein Domain Prediction)
GenTHREADER (Rapid Fold Recognition)
pDomTHREADER (Fold Domain Recognition)
Help...
Input Sequence (Single sequence or Multiple Sequence alignments; as raw sequence or fasta format)
Sequence
Help...
If you wish to test these services follow this link to retrieve a test fasta sequence.
Submission Details
Email Address for job completion alert (optional)
Help...
Password (only required for licenced commercial e-mail addresses)
Help...
Short identifier for submission
Solyc06g050510
Help...
Predict
Clear form

CHAPTER 11

Tertiary Structure Prediction

1. Frist find a structure similar to above sequence in PBD. We will use DELTA BLAST to search PBD.

2. Click on the first significant hit to access the PDB. In case, you don't have BLAST results, use the following link to access the previous results. Link

3. RCSB provides curated content of PDB and use PDB ID: 1Z3I to visualize the protein in RCSB.

4. Perform a multiple sequence alignment to find conserved sequences. |

:a:. Retrive sequence from databank

:b:. Selected sequences are in the following fasta file.

https://github.com/ajwije/2017_spring_Bioinfo_class/blob/master/rad54.fasta

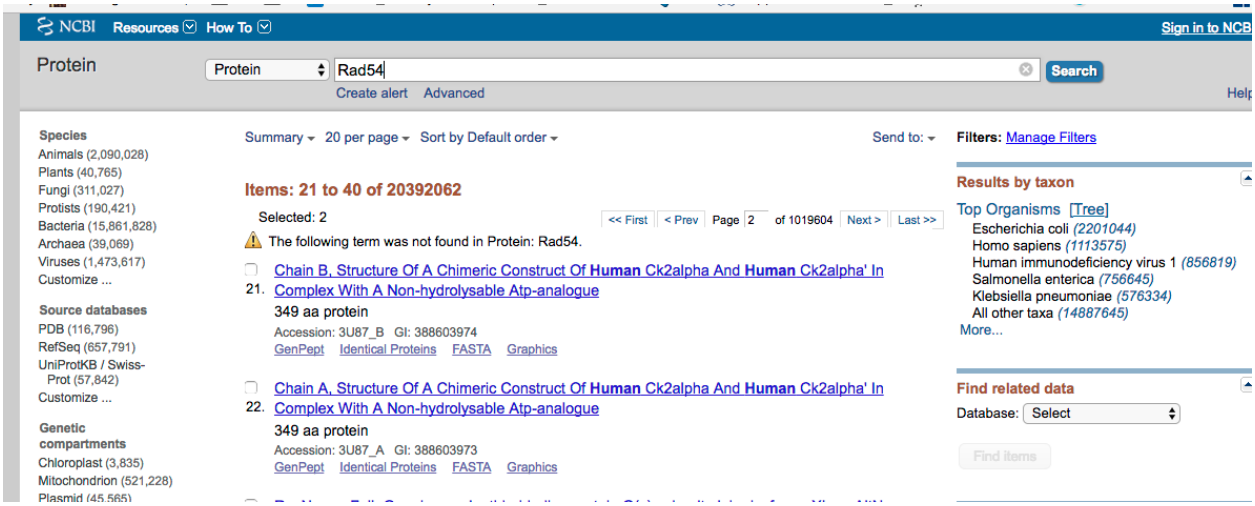
:c:. Use Tcoffee server and align the sequences using structural information:

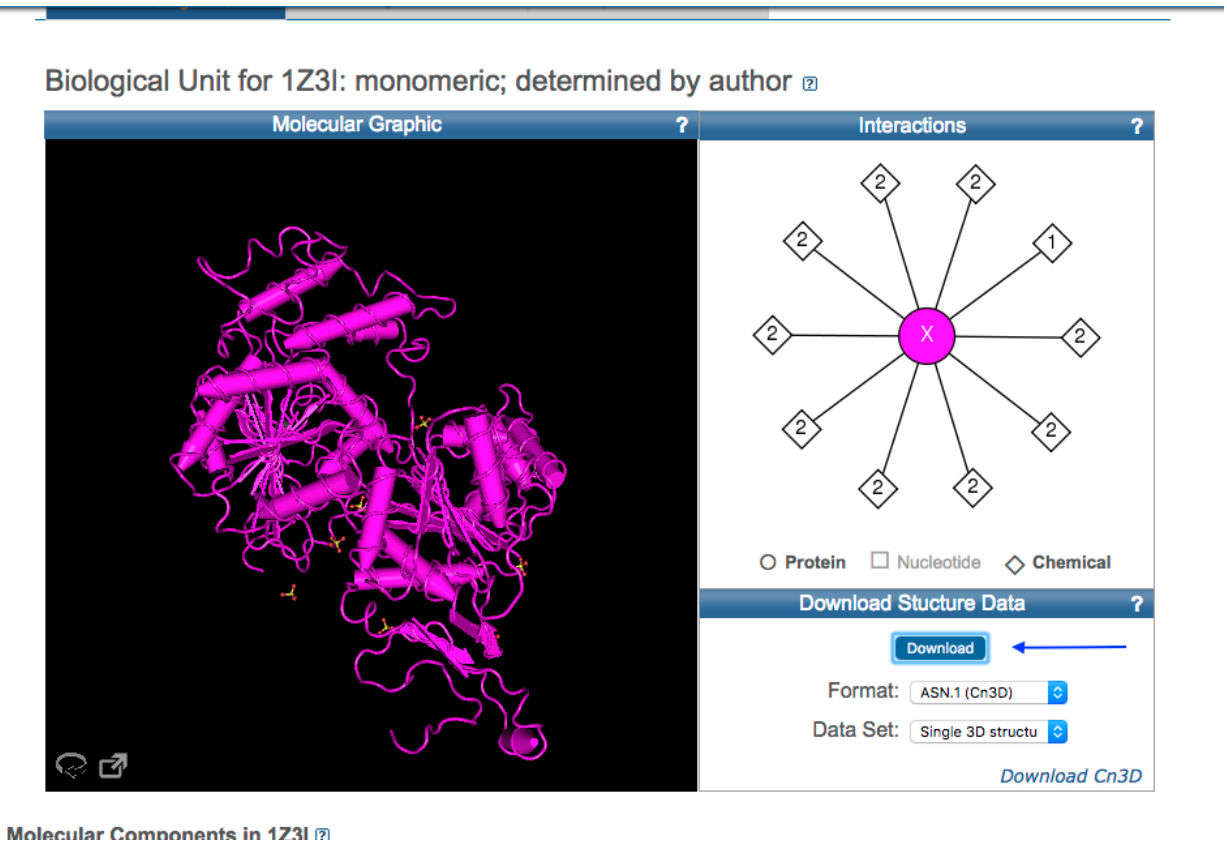
5. You can download crystal structure information from PDB in Cn3 format.

6. Download Cn3D software from NCBI and install it on your computer.

7. Open above Cn3 file using the Cn3D software.


blastn
blastp
blastx
tblastn
tblastx
BLASTP
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
GSKVKFVMSLIPRCLLRREKVLIFCHNIAPINLFLEIFERFYGWRKGIEVLVLQGDIELFQRGRIM
DLFEEP
GGPSKVMLASITTCAEGISLTAASRVILLDSEWNPSKSKQAIARAFRPGQDKVVYVYQLLATG
TLEEEKYKR
TTWKEWVSSMIFSEDLVEDPSHWQAPKIEDELLREIVEEDRATLFHAIMKNEKASNMGSLQE
Or, upload file
Choose File
No file chosen
Job Title
Protein Sequence (208 letters)
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Protein Data Bank proteins(pdb)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query
Optional
Create custom database
Enter an Entrez query to limit search
Program Selection
Algorithm
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database Protein Data Bank proteins(pdb) using DELTA-BLAST (Domain Enhanced Lookup


NCBI
Resources
How To
Sign in to NCB
Protein
Protein
Rad54
Search
Create alert
Advanced
Help
Species
Animals (2,090,028)
Plants (40,765)
Fungi (311,027)
Protists (190,421)
Bacteria (15,861,828)
Archaea (39,069)
Viruses (1,473,617)
Customize ...
Source databases
PDB (116,796)
RefSeq (657,791)
UniProtKB / Swiss-Prot (57,842)
Customize ...
Genetic compartments
Chloroplast (3,835)
Mitochondrion (521,228)
Summary
20 per page
Sort by Default order
Send to:
Filters: Manage Filters
Items: 21 to 40 of 20392062
Selected: 2
<< First
< Prev
Page 2 of 1019604
Next >
Last >>
The following term was not found in Protein: Rad54.
21. Chain B, Structure Of A Chimeric Construct Of Human Ck2alpha And Human Ck2alpha' In Complex With A Non-hydrolysable Atp-analogue
349 aa protein
Accession: 3U87_B GI: 388603974
GenPept Identical Proteins FASTA Graphics
22. Chain A, Structure Of A Chimeric Construct Of Human Ck2alpha And Human Ck2alpha' In Complex With A Non-hydrolysable Atp-analogue
349 aa protein
Accession: 3U87_A GI: 388603973
GenPept Identical Proteins FASTA Graphics
Results by taxon
Top Organisms [Tree]
Escherichia coli (2201044)
Homo sapiens (1113575)
Human immunodeficiency virus 1 (856819)
Salmonella enterica (756645)
Klebsiella pneumoniae (576334)
All other taxa (14887645)
More...
Find related data
Database: Select
Find items

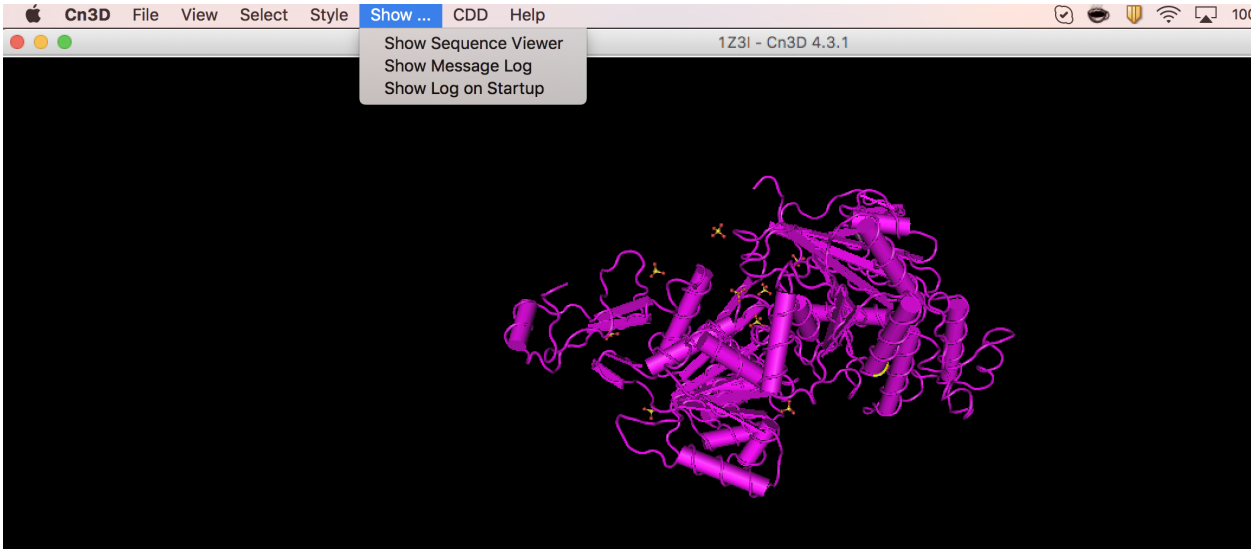


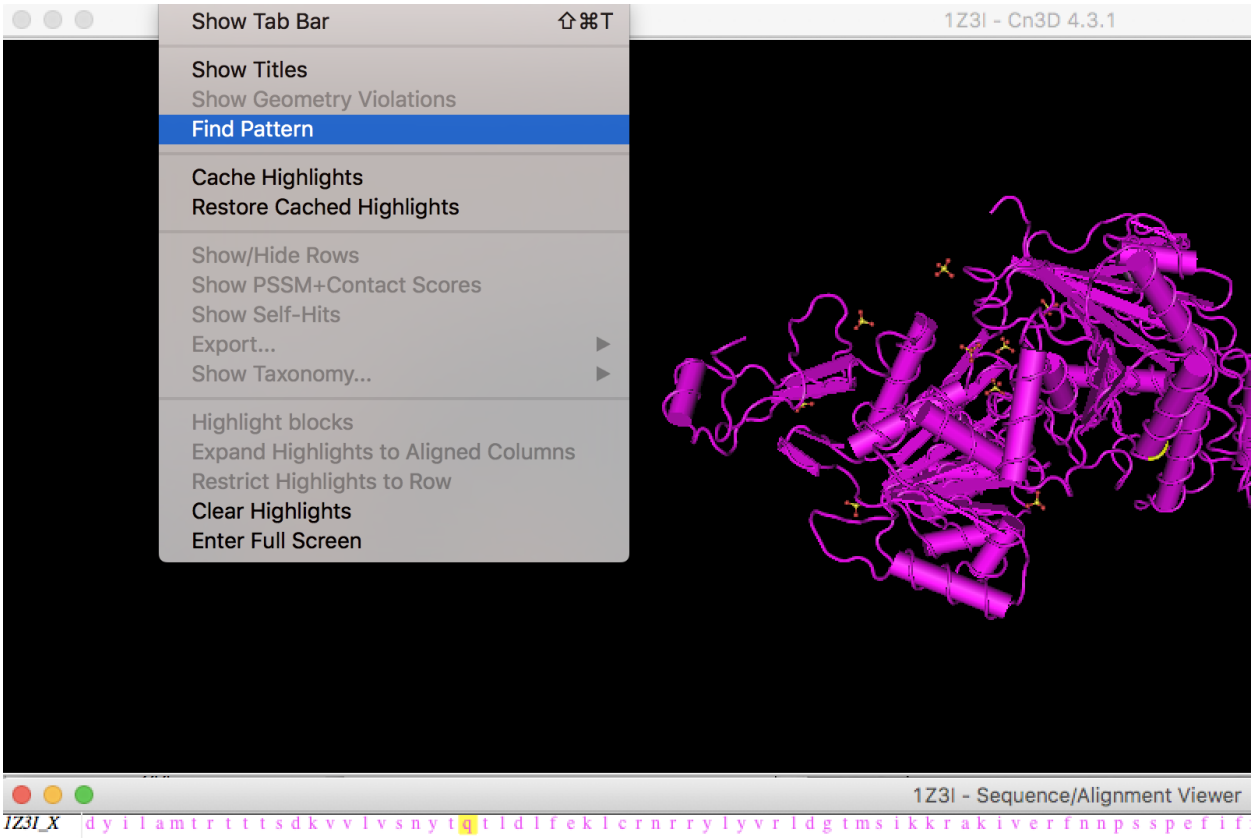
8. Go to sequence viewer

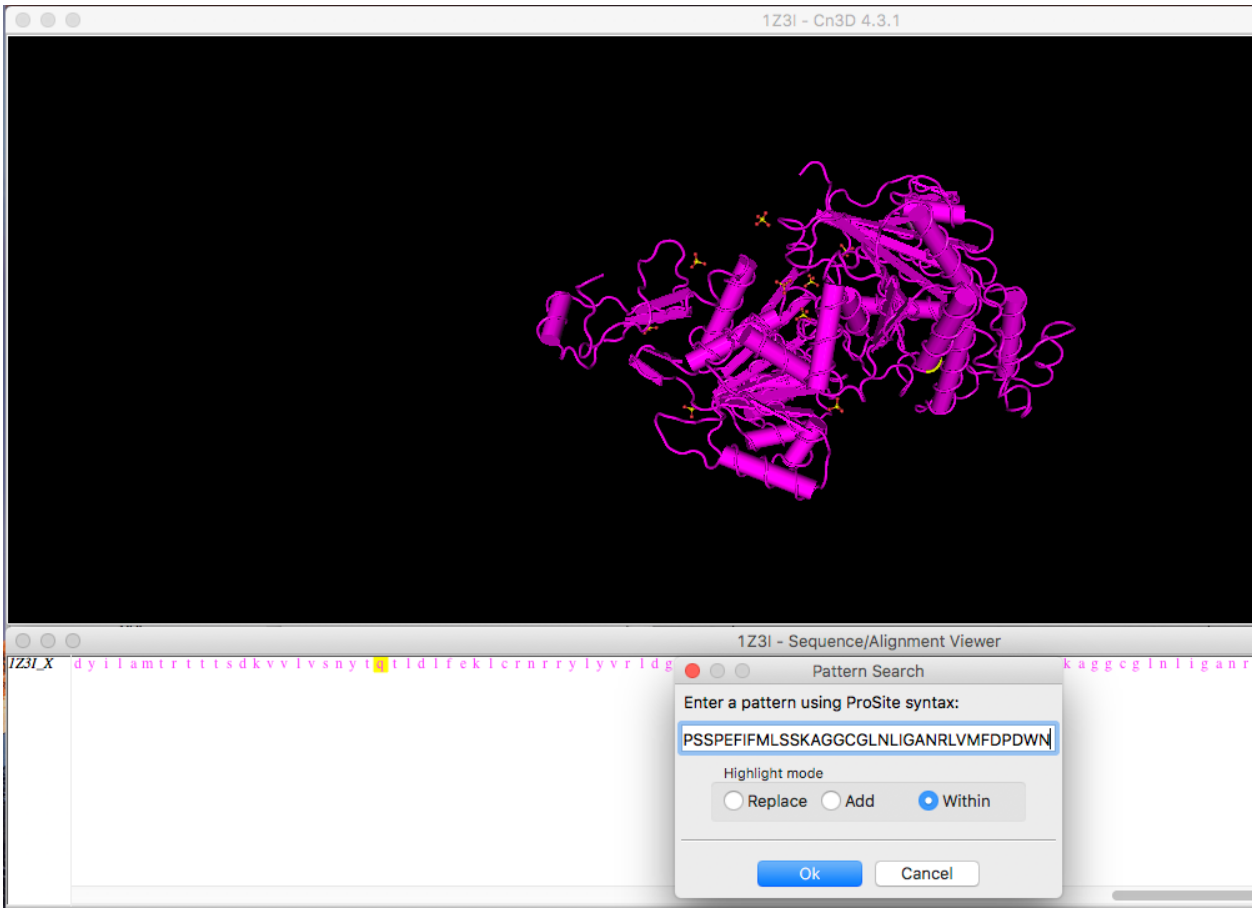
9. Under view, select find pattern:

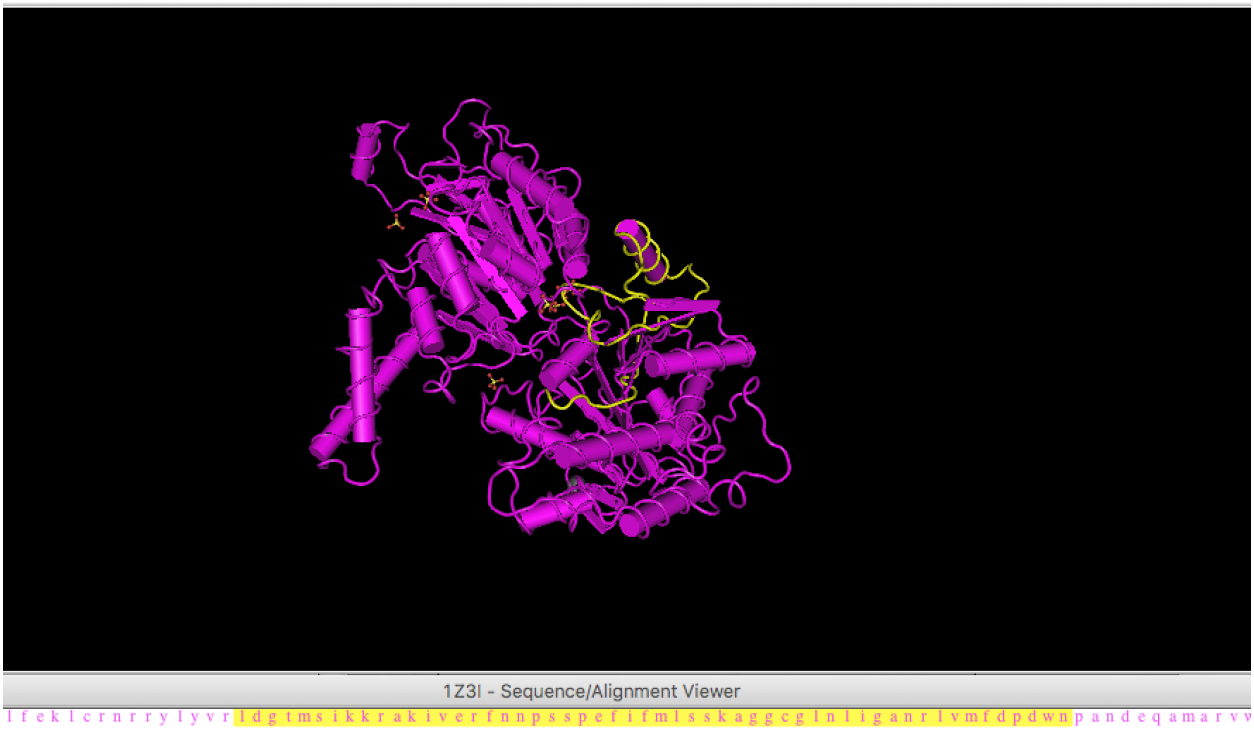
10. Copy a conserved region from multiple sequence alignment in the search window and click OK:

11. You will see conserved region displayed on the crystal structure.


Cn3D
File
View
Select
Style
Show ...
CDD
Help
Show Sequence Viewer
Show Message Log
Show Log on Startup
1Z3I - Cn3D 4.3.1


Show Tab Bar
⇧⌘T
Show Titles
Show Geometry Violations
Find Pattern
Cache Highlights
Restore Cached Highlights
Show/Hide Rows
Show PSSM+Contact Scores
Show Self-Hits
Export...
Show Taxonomy...
Highlight blocks
Expand Highlights to Aligned Columns
Restrict Highlights to Row
Clear Highlights
Enter Full Screen
1Z3I - Cn3D 4.3.1
1Z3I - Sequence/Alignment Viewer
1Z3I_X


1Z3I - Cn3D 4.3.1
1Z3I - Sequence/Alignment Viewer
1Z3I_X
Pattern Search
Enter a pattern using ProSite syntax:
PSSPEFIFMLSSKAGGCGLNLIGANRLVMFDPDWN
Highlight mode
Replace
Add
Within
Ok
Cancel


1Z3I - Sequence/Alignment Viewer
l f e k l c r n r r y l y v r l d g t m s i k k r a k i v e r f n n p s s p e f i f m l s s k a g g c g l n l i g a n r l v m f d p d w n p a n d e q a m a r v v

CHAPTER 12

The Delta-Delta Ct Method

Delta-Delta Ct method or Livak method is the most preferred method for qPCR data analysis. However, it can only be used when certain criteria are met. Please refer the lecture notes to make sure that these criteria are fulfilled. If not, more generalized method is called Pfaffl method. Please read the additional reading material to get more information about this method.

Here are the steps for Livak method:

The Excel file with all the calculation are in the qPCR analysis folder on Blackboard.

You have raw Ct (number of cycles that takes to reach threshold) for normal and tumor cells (3 replicates for each).

Samples	Raw Ct	
	GAPDH	p53
Tumor cells 1	21.00	23.00
Tumor cells 2	20.50	22.00
Tumor cells 3	20.60	22.50
Normal cells 1	20.00	26.00
Normal cells 2	20.50	26.20
Normal cells 3	20.30	26.40

12.1 Normalization

First, you will need calculate relative difference between the gene of interest (p53) and the house keeping gene (GAPDH).

Ct = Ct (gene of interest) – Ct (housekeeping gene)

Samples	Raw Ct		Delta Ct
	GAPDH	p53	
Tumor cells 1	21.00	23.00	=C3-B3
Tumor cells 2	20.50	22.00	
Tumor cells 3	20.60	22.50	
Normal cells 1	20.00	26.00	
Normal cells 2	20.50	26.20	
Normal cells 3	20.30	26.40	

12.2 Average of the control samples (normal cells)

As we compare our tumor (treatment) to control (normal cells), first we need to average the Ct for the 3 control (normal) samples.

12.3 Calculate the Ct relative to the average of Ct normal cells

Ct = Ct (Tumor sample) – Ct (normal average)

You can do this normal samples as well. Use $ signs infront of column number and raw letter (arrows) to fix the cell.

Samples	Raw Ct		Delta Ct	Del
	GAPDH	p53		
Tumor cells 1	21.00	23.00	2.00	
Tumor cells 2	20.50	22.00	1.50	
Tumor cells 3	20.60	22.50	1.90	
Normal cells 1	20.00	26.00	6.00	
Normal cells 2	20.50	26.20	5.70	
Normal cells 3	20.30	26.40	6.10	
Avg delta Ct			=average(E6:E8)	

Samples	Raw Ct		Delta Ct	Delta Delta ct	2
	GAPDH	p53			
Tumor cells 1	21.00	23.00	2.00	=E3-E9	
Tumor cells 2	20.50	22.00	1.50		
Tumor cells 3	20.60	22.50	1.90		
Normal cells 1	20.00	26.00	6.00		
Normal cells 2	20.50	26.20	5.70		
Normal cells 3	20.30	26.40	6.10		
Avg delta Ct			5.93		

12.4 Fold gene expression for each sample

Make sure you raise the negative Ct to power of two.

Fold gene expression = 2^-(Ct)

Samples	Raw Ct		Delta Ct	Delta Delta ct	2^delta d	A
	GAPDH	p53				
Tumor cells 1	21.00	23.00	2.00	-3.93	=2^-(F3)	
Tumor cells 2	20.50	22.00	1.50	-4.43		
Tumor cells 3	20.60	22.50	1.90	-4.03		
Normal cells 1	20.00	26.00	6.00	0.07		
Normal cells 2	20.50	26.20	5.70	-0.23		
Normal cells 3	20.30	26.40	6.10	0.17		
Avg delta Ct			5.93			

12.5 Overall fold change

You can calculate average fold change for both tumor and normal samples. Ratio between these two the fold change between tumor and normal samples.

Enter

B	C	D	E	F	G	H
Samples	Raw Ct		Delta Ct	Delta Delta ct	2^delta delta Ct	Log10
	GAPDH	p53				
Tumor cells 1	21.00	23.00	2.00	-3.93	15.27746566	
Tumor cells 2	20.50	22.00	1.50	-4.43	21.60559914	
Tumor cells 3	20.60	22.50	1.90	-4.03	16.37398227	
Normal cells 1	20.00	26.00	6.00	0.07	0.954841604	
Normal cells 2	20.50	26.20	5.70	-0.23	1.175547906	
Normal cells 3	20.30	26.40	6.10	0.17	0.890898718	
Avg delta Ct			5.93			
Average Tumor cells	=average(G3:G5)					
Average Normal cells						

Samples	Raw Ct		Delta Ct	Delta Delta ct	2^delta delta Ct	Log10
	GAPDH	p53				
Tumor cells 1	21.00	23.00	2.00	-3.93	15.27746566	
Tumor cells 2	20.50	22.00	1.50	-4.43	21.60559914	
Tumor cells 3	20.60	22.50	1.90	-4.03	16.37398227	
Normal cells 1	20.00	26.00	6.00	0.07	0.954841604	
Normal cells 2	20.50	26.20	5.70	-0.23	1.175547906	
Normal cells 3	20.30	26.40	6.10	0.17	0.890898718	
Avg delta Ct			5.93			
Average Tumor cells	17.752349					
Average Normal cells	=average(G6:G8)					

Samples	Raw Ct		Delta Ct	Delta Delta ct	2^delta delta Ct	Log10
	GAPDH	p53				
Tumor cells 1	21.00	23.00	2.00	-3.93	15.27746566	
Tumor cells 2	20.50	22.00	1.50	-4.43	21.60559914	
Tumor cells 3	20.60	22.50	1.90	-4.03	16.37398227	
Normal cells 1	20.00	26.00	6.00	0.07	0.954841604	
Normal cells 2	20.50	26.20	5.70	-0.23	1.175547906	
Normal cells 3	20.30	26.40	6.10	0.17	0.890898718	
Avg delta Ct			5.93			
Average Tumor cells	17.7523					
Average Normal cells	1.0071					
Fold change Tumor/normal	=C11/C12					

12.6 Log transformation

To perform parametric statistical tests such as T-test, it advised to transform the final gene expression results to log values (any log base). This would make data distribution symmetric.

Here we have change the 2^-(Ct) to log 10.

Samples	Raw Ct		Delta Ct	Delta Delta ct	2^delta delta Ct	Log10
	GAPDH	p53				
Tumor cells 1	21.00	23.00	2.00	-3.93	15.27746566	1.184051316
Tumor cells 2	20.50	22.00	1.50	-4.43	21.60559914	1.334566314
Tumor cells 3	20.60	22.50	1.90	-4.03	16.37398227	1.214154316
Normal cells 1	20.00	26.00	6.00	0.07	0.954841604	-0.020068666
Normal cells 2	20.50	26.20	5.70	-0.23	1.175547906	0.070240332
Normal cells 3	20.30	26.40	6.10	0.17	0.890898718	-0.050171666
Avg delta Ct			5.93			
Average Tumor cells	17.75234902					
Average Normal cells	1.007096076					
Tumor SEM	1.952454822					
Normal SEM	0.086224876					
Fold change Tumor/normal	17.62726461					
T-TEST	=TTEST(H3:H5,H6:H8, 2,3)					

Two tail test

Unequal variance

12.7 T-test

Need to be careful when using parametric tests if data is not normally distributed, it would lead to erroneous conclusions.

Samples	Raw Ct		Delta Ct	Delta Delta ct	2^delta delta Ct	Log10
	GAPDH	p53				
Tumor cells 1	21.00	23.00	2.00	-3.93	15.27746566	1.184051316
Tumor cells 2	20.50	22.00	1.50	-4.43	21.60559914	1.334566314
Tumor cells 3	20.60	22.50	1.90	-4.03	16.37398227	1.214154316
Normal cells 1	20.00	26.00	6.00	0.07	0.954841604	-0.020068666
Normal cells 2	20.50	26.20	5.70	-0.23	1.175547906	0.070240332
Normal cells 3	20.30	26.40	6.10	0.17	0.890898718	-0.050171666
Avg delta Ct			5.93			
Average Tumor cells	17.75234902					
Average Normal cells	1.007096076					
Tumor SEM	1.952454822					
Normal SEM	0.086224876					
Fold change Tumor/normal	17.62726461					
T-TEST	=TTEST(H3:H5,H6:H8, 2,3)					

Two tail test

Unequal variance

Select log 10 of 2^-(Ct) values for Normal and tumor samples as indicated. Use two tail test (number 2) and assuming unequal variance (3).

Resulting P value is less than 0.05 and therefore, we reject the null hypothesis and two sample means are significantly different at 0.05 level.

Samples	Raw Ct		Delta Ct	Delta Delta ct	2^delta delta Ct	Log10
	GAPDH	p53				
Tumor cells 1	21.00	23.00	2.00	-3.93	15.27746566	1.184051316
Tumor cells 2	20.50	22.00	1.50	-4.43	21.60559914	1.334566314
Tumor cells 3	20.60	22.50	1.90	-4.03	16.37398227	1.214154316
Normal cells 1	20.00	26.00	6.00	0.07	0.954841604	-0.020068666
Normal cells 2	20.50	26.20	5.70	-0.23	1.175547906	0.070240332
Normal cells 3	20.30	26.40	6.10	0.17	0.890898718	-0.050171666
Avg delta Ct			5.93			
Average Tumor cells	17.75234902					
Average Normal cells	1.007096076					
Fold change Tumor/normal	17.62726461					
T-TEST	4.39604E-05					

P-value

. module:: Getting Data into Galaxy

synopsis

CHAPTER 13

Getting data into Galaxy

13.1 Step 1: Login into Galaxy

Click on the following link to go to the ASU Galaxy site:

https://orpheus.cs.astate.edu/

Use your ASU username and password to login to Galaxy.

13.2 Step 2: Getting data

Data we are using for this analysis came from Loraine et al, 2015 study. In the original study, there are 10 samples (Five Controls and heat treated). Here we are using only 3 samples for each group (3 control and 3 heat treated). These files were downloaded from NCBI's Short Read Archive (SRA) using SRA toolkit.

Use the following links to get data. Each link is one data file.

https://de.cyverse.org/dl/d/2AB5824F-73BA-4C6B-8530-457609F632BA/Control2_1.fastq

https://de.cyverse.org/dl/d/46E690E4-C4A0-495F-9B11-F12AD9A25EE3/Control2_2.fastq

https://de.cyverse.org/dl/d/7FEE6359-24AE-478D-A0B1-C6D2CA09E45E/Control3_1.fastq

https://de.cyverse.org/dl/d/8FBB264D-F0CA-4F2C-821A-DB1C709315B2/Control3_2.fastq

https://de.cyverse.org/dl/d/9A45E994-2CDC-4643-AC4C-45C9625138F6/Heat1_1.fastq

https://de.cyverse.org/dl/d/46093383-493A-4D4E-A607-D3E56916DF59/Heat1_2.fastq

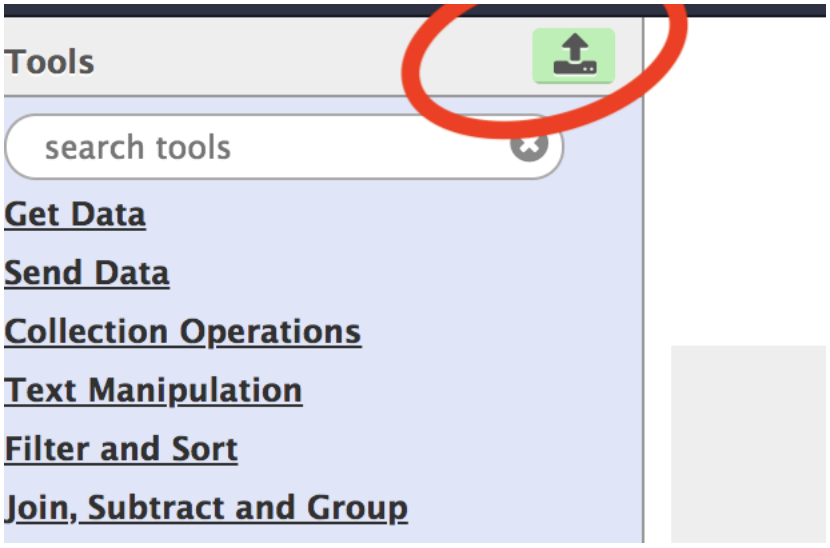
https://de.cyverse.org/dl/d/9668B243-7009-4AD3-BBDA-350D6A60119D/Heat2_1.fastq

https://de.cyverse.org/dl/d/FE1C3CC3-9133-4244-BCBB-816B8D2D5F97/Heat2_2.fastq

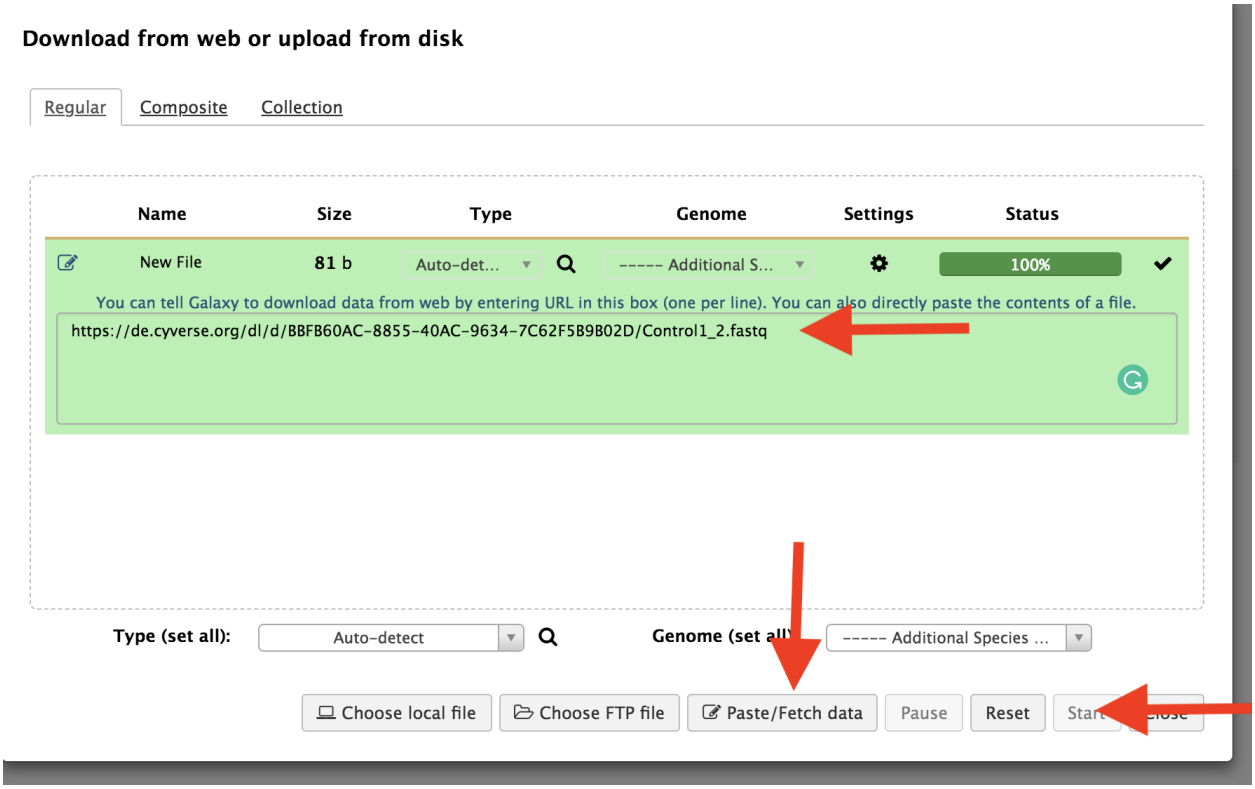
https://de.cyverse.org/dl/d/D635B6EE-BE26-4BC4-A058-3E51B1AA69C4/Heat3_1.fastq

https://de.cyverse.org/dl/d/F88561AF-CFF2-4FC8-B6B4-D8623779BB24/Heat3_2.fastq

13.3 Step 1: Click on the upload icon on upper left hand corner



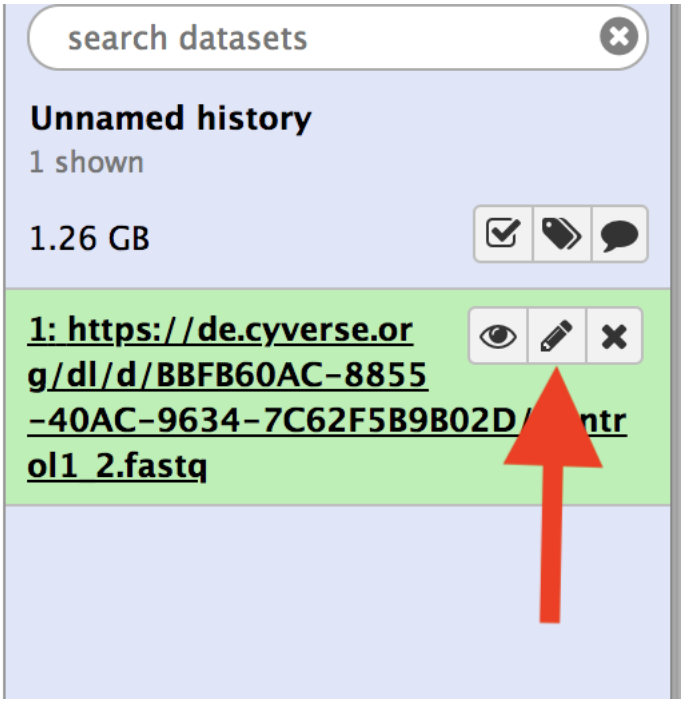
13.4 Step 2: Copy one of the links above. Click on the Paste/Fetch icon and paste link in the box. Click on start.



13.5 Step 3: One the data is uploaded, they will appear in the right hand panel. You can use the pencil icon to change the name.

13.6 Reference:

Loraine AE, Blakley IC, Jagadeesan S, Harper J, Miller G, Firon N. Analysis and Visualization of RNA-Seq Expression Data Using RStudio, Bioconductor, and Integrated Genome Browser. *Methods Mol Biol.* 2015;1284:481-501. doi: 10.1007/978-1-4939-2444-8_24. PubMed PMID: 25757788.

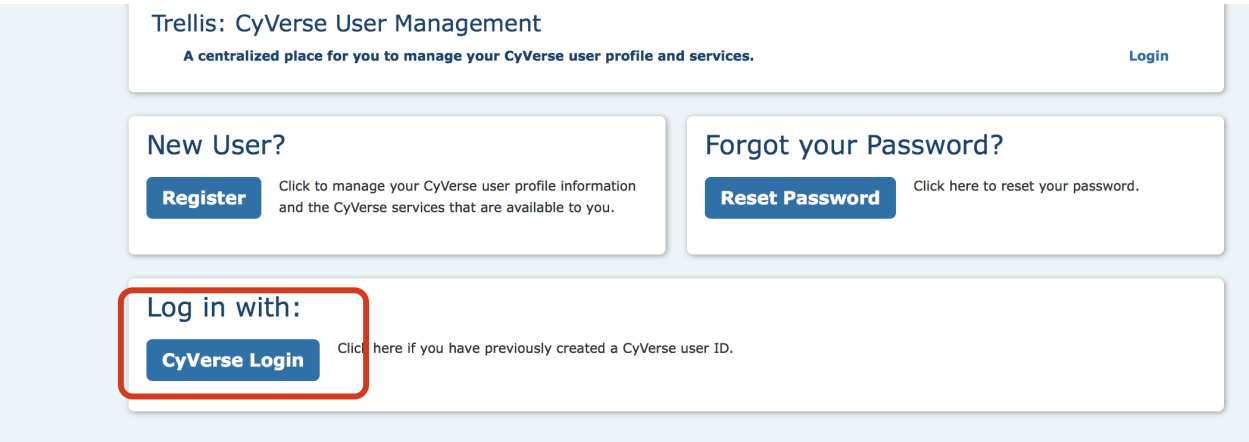

search datasets
Unnamed history
1 shown
1.26 GB
1: https://de.cyverse.org/dl/d/BBFB60AC-8855-40AC-9634-7C62F5B9B02D/
ntrol1_2.fastq

CHAPTER 14

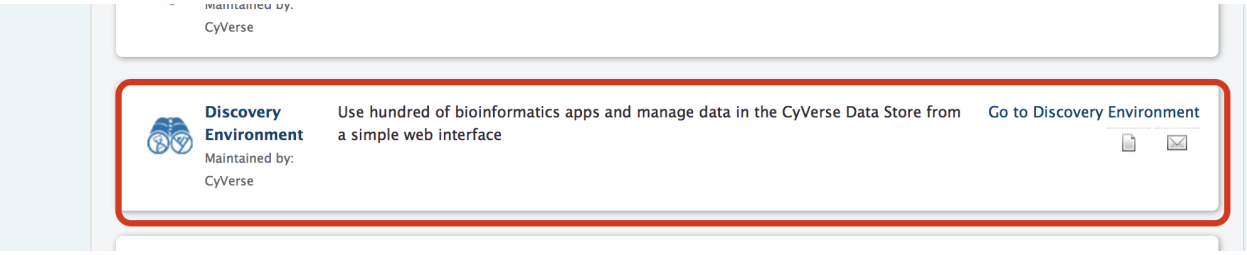
FastQC analysis using Galaxy

14.1 Step 1: Login into Galaxy

First login to your Cyverse account using your name and password.

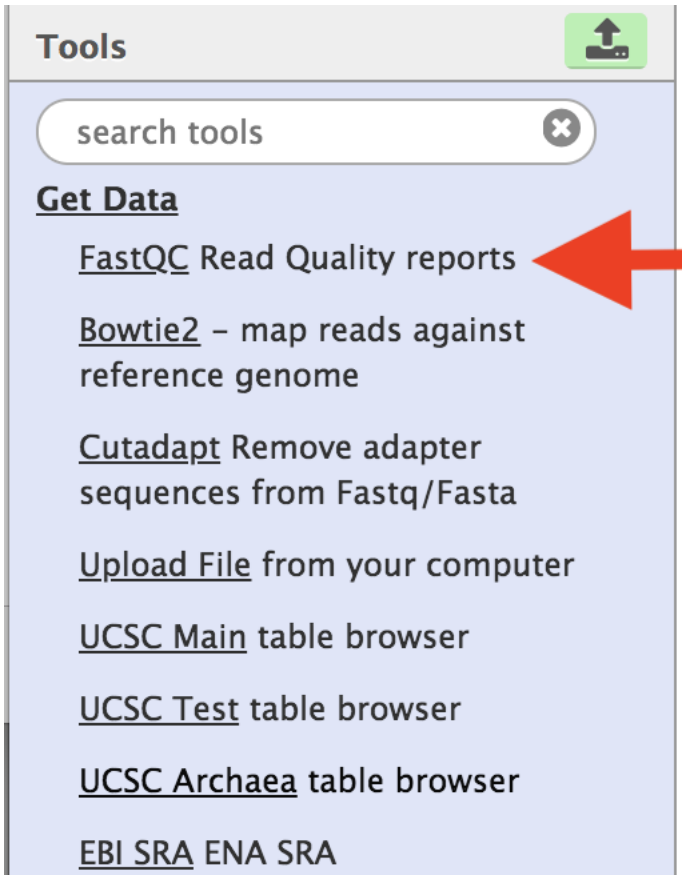


Then, go to your DE account.

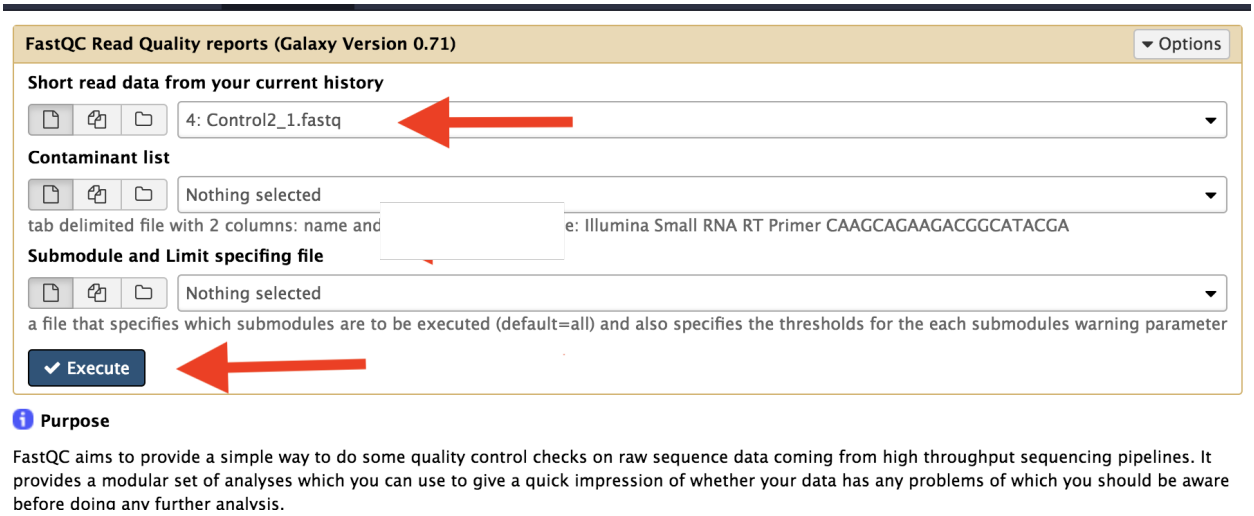


14.2 Step 3: Performing FastQC analysis:

First step of the data analysis is to check the quality of the sequences. For this purpose, we are using the FastQC tool on Galaxy. a. Type FastQC in the search box on top left hand corner.



d. Select a fastq file and execute the analysis.



CHAPTER 15

Adapter and quality trimming using Cutadapt

We are going to use Trim-galore to trim adapters, and poor quality bases. This tool has several advantages. It allows selection multiple files. You can also select both forward and reverse reads. If you want to read more about Trim-galore, please visit their website. Also, Trim-galore is a wrapper for Cutadapt , which is the actual tool that performs the trimming.

Please follow the tutorial carefully.

15.1 Step 1: Launching Cutadapt and performing the analysis

1. Type Trim-galore in the search box on top left hand corner.

2. Select paired-end and select the two paired-end files are shown below. Use Illumina universal adapter to trim.

3. Click on the advance settings. Set the parameters as indicated in blue arrows.

Cutadapt Remove adapter sequences from Fastq/Fasta (Galaxy Version 1.16.3) Options

Single-end or Paired-end reads?

Paired-end

FASTQ/A file #1

125: Heat3_1.fastq

Should be of datatype "fastq.gz"or "fasta"

FASTQ/A file #2

126: Heat3_2.fastq

Should be of datatype "fastq.gz"or "fasta"

Read 1 Options

3' (End) Adapters

1: 3' (End) Adapters

Source

Enter custom sequence

Enter custom 3' adapter name (Optional)

Enter custom 3' adapter sequence

AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC

(-a)

Read 2 Options

3' (End) Adapters

1: 3' (End) Adapters

Source

Enter custom sequence

Enter custom 3' adapter name (Optional)

Enter custom 3' adapter sequence

AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT

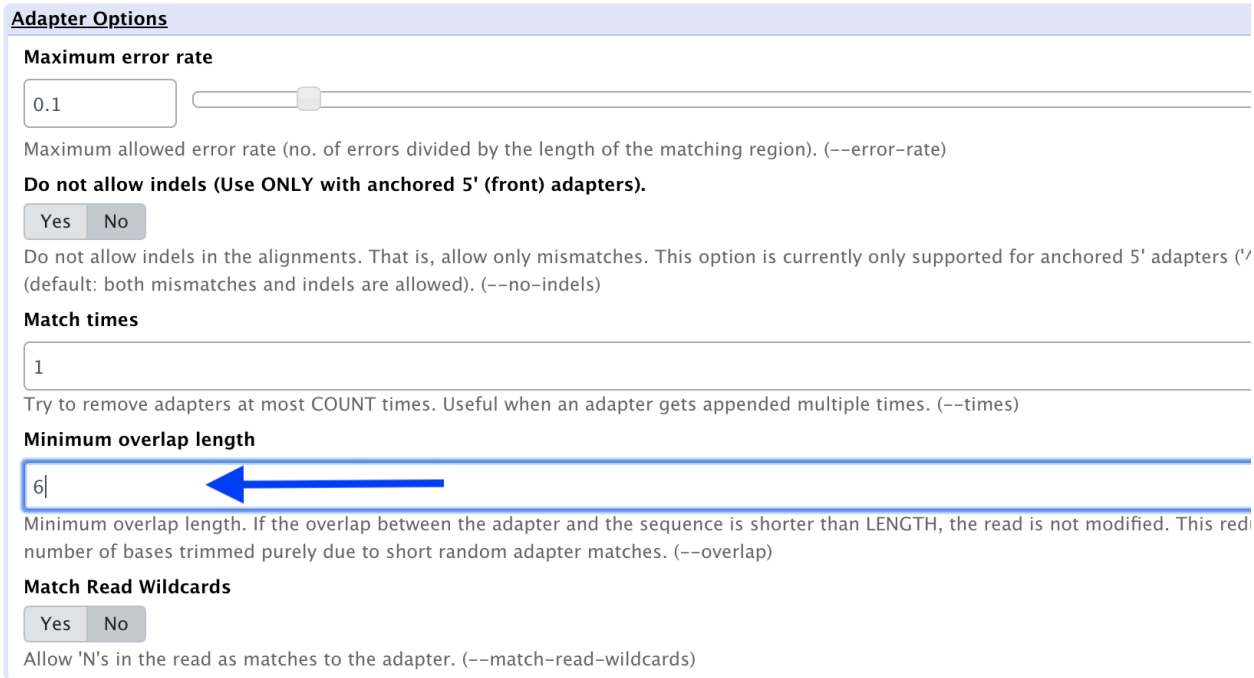
(-A)

Insert 3' (End) Adapters

Sequence of an adapter ligated to the 3' end of the second read in each pair. The adapter and subsequent bases are trimmed. If a '$' character is appended ('anchoring'), the adapter is only found if it is a suffix of the read. To search for a linked adapter, separate the 2 sequences with 3 dots (ADAPTER1...ADAPTER2), see Help below.

5' (Front) Adapters

Insert 5' (Front) Adapters



4. Launch the analysis.

Filter Options

Discard Trimmed Reads

Yes No

Discard reads that contain the adapter instead of trimming them. Use the 'Minimum overlap length' option in ord many randomly matching reads! (--discard-trimmed)

Discard Untrimmed Reads

Yes No

Discard reads that do not contain the adapter. (--discard_untrimmed)

Minimum length

40

Discard trimmed reads that are shorter than LENGTH. Reads that are too short even before adapter removal are a initial primer is not counted. Value of 0 means no minimum length. (--minimum-length)

Maximum length

0

Discard trimmed reads that are longer than LENGTH. Reads that are too long even before adapter removal are als initial primer is not counted. Value of 0 means no maximum length. (--maximum-length)

Read Modification Options

Quality cutoff

20

Trim low-quality bases from 5' and/or 3' ends of each read before adapter remo
only the 3' end is trimmed. If two comma-separated cutoffs are given, the 5' end
quality-cutoff)

NextSeq trimming

0

Experimental option for quality trimming of NextSeq data. This is necessary beca
end of the fragment (it encodes G as 'black'). This option works like regular qual
qualities of G bases are ignored. (--nextseq-trim)

Trim Ns

Yes No

Trim N's on ends of reads. (--trim-n)

Prefix

Add this prefix to read names (--prefix)

Suffix

Add this suffix to read names (--suffix)

Strip suffix

Remove this suffix from read names if present. (--strip-suffix)

Length

40

CHAPTER 16

Adapter and quality trimming using trim-galore

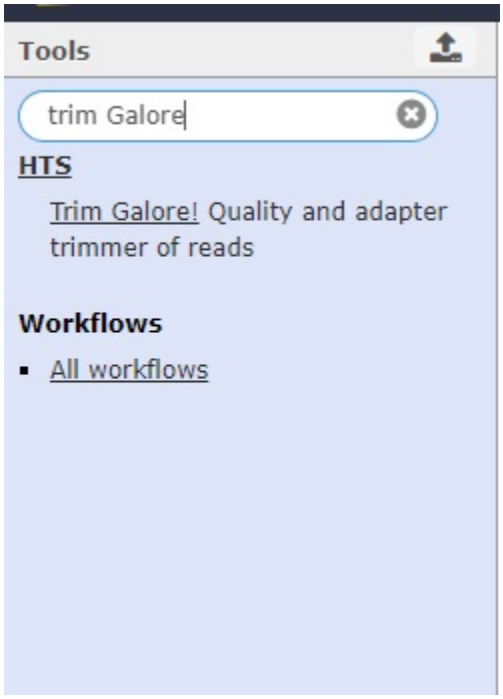
We are going to use Trim-galore to trim adapters, and poor quality bases. This tool has several advantages. It allows selection multiple files. You can also select both forward and reverse reads. If you want to read more about Trim-galore, please visit their website. Also, Trim-galore is a wrapper for Cutadapt , which is the actual tool that performs the trimming.

Please follow the tutorial carefully.

16.1 Step 1: Launching Trim-galore

1. In the finder window type "trim-galore"

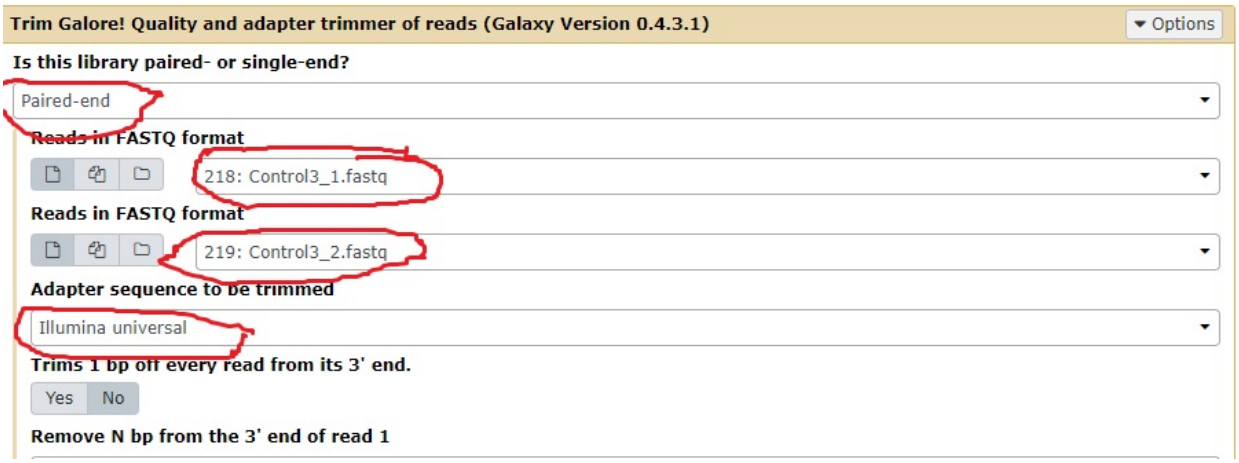
3. Select "Trim Galore".



16.2 Step 2: Selecting input files

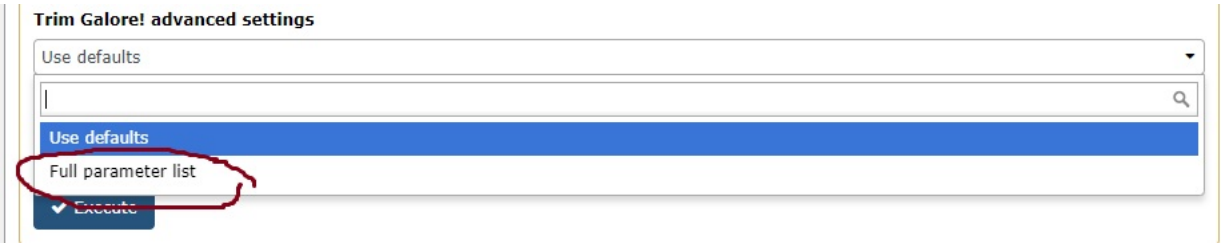
As indicated in the figure: 1. Is this library paired- or single-end? "Paired-end"

2. Adapter sequence to be trimmed: "Select Illumina universal"

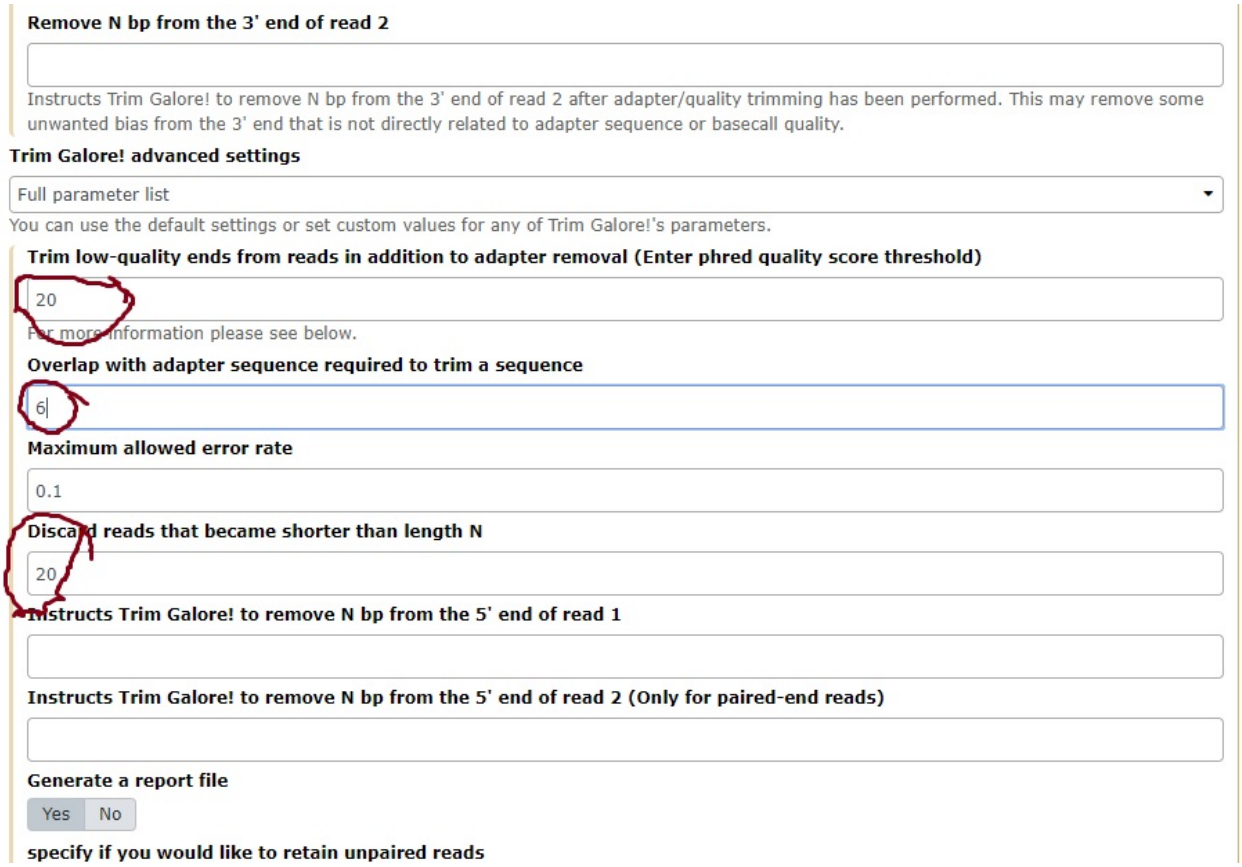


16.3 Step 3: Advance settings

1. From “Trim Galore! advanced settings” select “Full parameter list”



2. Set “Overlap with adapter sequence required to trim a sequence” to 6.



3. Run the analysis.

CHAPTER 17

Use Splice aware aligner, Tophat2 to align short reads

1. We are using Paired end reads and set the "Is this library mate-paired?" pulldown to "Paired-end", then a second pulldown will appear to specify the 2nd FASTQ.
2. "Mean Inner Distance between Mate Pairs" value for this parameter should obtained from the person incharge of the sequencing.
3. Mean Inner Distance between Mate Pairs = length of the Fragments used for sequencing – (Length of Illumina adapters (often 120bp) + part sequenced (76+76))
4. Genome should be obtained from the SolGenome.net (ftp://ftp.solgenomics.net/tomato_genome/assembly/build_3.00/) and select it from the history.
5. This library has been prepared to preserve the strandedness of the RNAs.
6. Minimum and maximum intron lengths should be changed according to genome used.
7. Change the intron lengths for split reads as well.

TopHat Gapped-read mapper for RNA-seq data (Galaxy Version 2.1.1) Options

Is this single-end or paired-end data?

Paired-end (as individual datasets)

RNA-Seq FASTQ file, forward reads

4: Control2_1.fastq

Must have Sanger-scaled quality values with ASCII offset 33

RNA-Seq FASTQ file, reverse reads

5: Control2_2.fastq

Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs

200

-r/--mate-inner-dist; This is the expected (mean) inner distance between mate pairs. For, example, for paired end runs with fragments selected at 300bp, where each end is 50bp, you should set -r to be 200. The default is 50bp.

Std. Dev for Distance between Mate Pairs

40

--mate-std-dev; The standard deviation for the distribution on inner distances between mate pairs. The default is 20bp.

Report discordant pair alignments?

Yes

--no-discordant

Use a built in reference genome or own from your history

Use a genome from history

Built-ins genomes were created using default options

Select the reference genome

24: S_lycopersicum_chromosomes.3.00.fa

TopHat settings to use

TopHat settings to use

Full parameter list

You can use the default settings or set custom values for any of Tophat's parameters.

Max realign edit distance

1000

--read-realign-edit-dist; Some of the reads spanning multiple exons may be mapped incorrectly as a contiguous alignment to the genome even though the correct alignment should be a spliced one - this can happen in the presence of processed pseudogenes that are rarely (if at all) transcribed or expressed. This option can direct TopHat to re-align reads for which the edit distance of an alignment obtained in a previous mapping step is above or equal to this option value. If you set this option to 0, TopHat will map every read in all the mapping steps (transcriptome if you provided gene annotations, genome, and finally splice variants detected by TopHat), reporting the best possible alignment found in any of these mapping steps. This may greatly increase the mapping accuracy at the expense of an increase in running time. The default value for this option is set such that TopHat will not try to realign reads already mapped in earlier steps.

Max edit distance

2

--read-edit-dist; Final read alignments having more than these many edit distance are discarded.

Library Type

FR Unstranded

--library-type; TopHat will treat the reads as strand specific. Every read alignment will have an XS attribute tag. Consider supplying library type options below to select the correct RNA-seq protocol.

Final read mismatches

2

--read-mismatches; Final read alignments having more than these many mismatches are discarded.

Use bowtie -n mode

No

--bowtie-n; TopHat uses "-v" in Bowtie for initial read mapping (the default), but with this option, "-n" is used instead. Read segments are always mapped using "-v" option.

Anchor length (at least 3)

8

-a/--min-anchor-length; TopHat will report junctions spanned by reads with at least this many bases on each side of the junction. Note that individual spliced alignments may span a junction with fewer than this many bases on one side. However, every junction involved in spliced alignments is supported by at least one read with this many bases on each side. This must be at least 3 and the default is 8.

Maximum number of mismatches that can appear in the anchor region of spliced alignment

0

-m/--splice-mismatches; The default is 0.

The minimum intron length

70

-i/--min-intron-length; TopHat will ignore donor/acceptor pairs closer than this many bases apart. The default is 70.

The maximum intron length

20,000

-I/--max-intron-length; When searching for junctions ab initio, TopHat will ignore donor/acceptor pairs farther than this many bases apart, except when such a pair is supported by a split segment alignment of a long read. The default is 500000.

Allow indel search

Yes

Max insertion length.

3

--max-insertion-length; The maximum insertion length. The default is 3.

Max deletion length.

3

--max-deletion-length; The maximum deletion length. The default is 3.

Maximum number of alignments to be allowed

20

-g/--max-multihits; Instructs TopHat to allow up to this many alignments to the reference for a given read, and choose the alignments based on their alignment scores if there are more than this number. The default is 20 for read mapping. Unless you use --report-secondary-alignments,

Maximum number of alignments to be allowed

20

-g/--max-multihits; Instructs TopHat to allow up to this many alignments to the reference for a given read, and choose the alignments based on their alignment scores if there are more than this number. The default is 20 for read mapping. Unless you use --report-secondary-alignments, TopHat will report the alignments with the best alignment score. If there are more alignments with the same score than this number, TopHat will randomly report only this many alignments. In case of using --report-secondary-alignments, TopHat will try to report alignments up to this option value, and TopHat may randomly output some of the alignments with the same score to meet this number.

Minimum intron length that may be found during split-segment (default) search

50

--min-segment-intron; The minimum intron length that may be found during split-segment search. The default is 50.

Maximum intron length that may be found during split-segment (default) search

20000

--max-segment-intron; The maximum intron length that may be found during split-segment search. The default is 500000.

Number of mismatches allowed in each segment alignment for reads mapped independently

2

--segment-mismatches; Read segments are mapped independently, allowing up to this many mismatches in each segment alignment. The default is 2.

Minimum length of read segments

25

--segment-length; Each read is cut up into segments, each at least this long. These segments are mapped independently. The default is 25.

Output unmapped reads

Yes No

If checked, a BAM with the unmapped reads will be added to the history

Do you want to supply your own junction data

No

The options below allow you validate your own list of known transcripts or junctions with your RNA-Seq data. Note that the chromosome names in the files provided with the options below must match the names in the Bowtie index.

17.1 Output files:

1. accepted_hits (BAM, BAI)

2. Two binary files: .BAM (data) and .BAI (index)

3. These are the actual paired reads mapped to their position on the genome, and split across exon junctions. This can be visualized in IGV, IGB or UCSC, but you must download both .BAM and .BAI files to the same directory. splice_junctions (BED)

4. BED file (list of genomic locations, no sequence) listing all the places TopHat had to split a read into two pieces to span an exon junction. This can be visualized at UCSC or in IGV, etc.

5. deletions (BED) (if indel search is on)

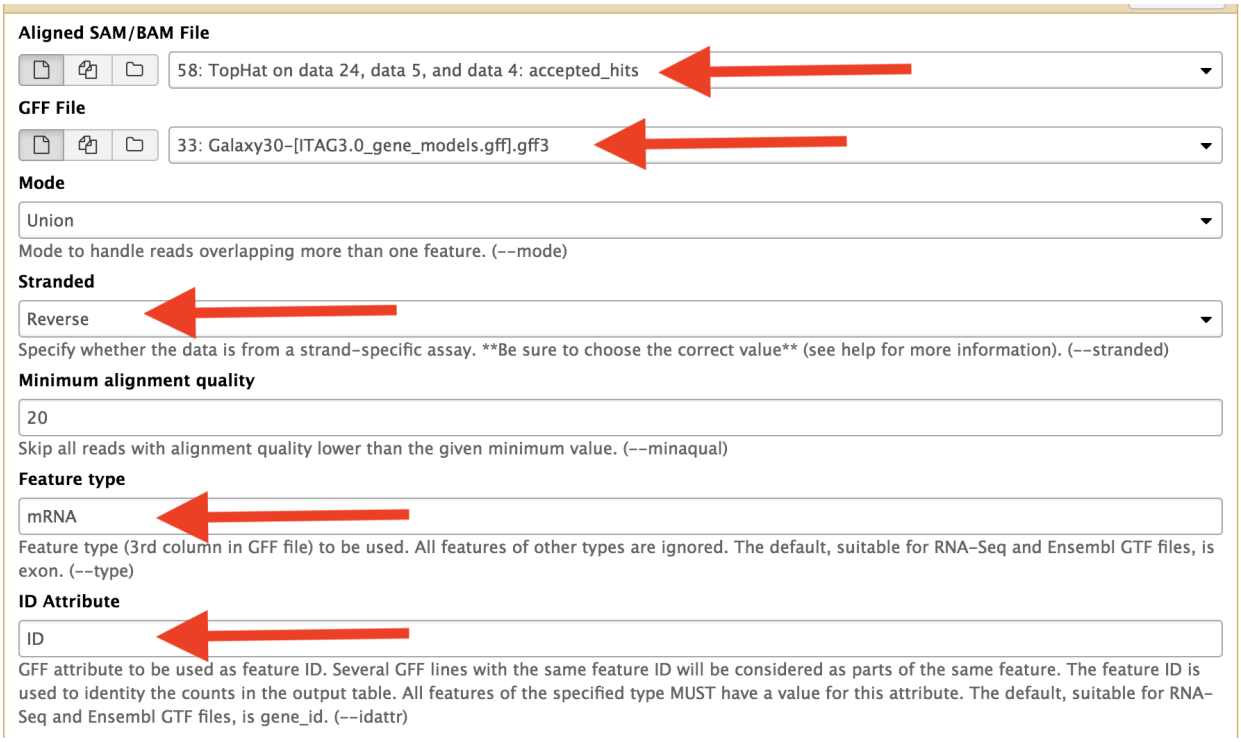
6. insertions (BED) (if indel search is on)

CHAPTER 18

Use Htseq to counts reads mapped to features

Use Htseq to counts the reads aligned to exons on the genes. Change the parameters as indicated in red arrows.

7. Change the intron lengths for split reads as well.


Aligned SAM/BAM File
58: TopHat on data 24, data 5, and data 4: accepted_hits
GFF File
33: Galaxy30-[ITAG3.0_gene_models.gff].gff3
Mode
Union
Mode to handle reads overlapping more than one feature. (--mode)
Stranded
Reverse
Specify whether the data is from a strand-specific assay. **Be sure to choose the correct value** (see help for more information). (--stranded)
Minimum alignment quality
20
Skip all reads with alignment quality lower than the given minimum value. (--minaqual)
Feature type
mRNA
Feature type (3rd column in GFF file) to be used. All features of other types are ignored. The default, suitable for RNA-Seq and Ensembl GTF files, is exon. (--type)
ID Attribute
ID
GFF attribute to be used as feature ID. Several GFF lines with the same feature ID will be considered as parts of the same feature. The feature ID is used to identity the counts in the output table. All features of the specified type MUST have a value for this attribute. The default, suitable for RNA-Seq and Ensembl GTF files, is gene_id. (--idattr)

CHAPTER 19

Use Kellisto to map reads to cDNA and count

Kellisto is an ultrafast alignment-free quantification tool.

1. Change parameters as indicated.

2. cDNA file can be obtained from the Solgenome.net(ftp://ftp.solgenomics.net/tomato_genome/annotation/ITAG3.0_release/ITAG3.0_cDNA.fasta) |

3. Once you get the results, click on the "eye" icon on the history pane. Then click on the "disk" icon on the left bottom left-hand corner to download the data into your computer.

4. Open the download file with Excel.

5. In Excel, click on "Data" and then click on "Text to Column". Separate column using tab to separate data into columns. Sum the numbers in "est_counts" using Auto Sum function.


Reference transcriptome for quantification
Use a transcriptome from history
FASTA reference transcriptome
43: ITAG3.0_cDNA.fasta
Single-end or paired reads
Paired
Collection or individual datasets
Individual files
Forward reads
4: Control2_1.fastq
Reverse reads
5: Control2_2.fastq
Perform sequence based bias correction
Yes
No
(--bias)
Number of bootstrap samples
10
default: 0 (--bootstrap-samples)
Seed for the bootstrap sampling
42
default: 42 (--seed)
Search for fusions
Yes
No
for Pizzly (--fusion)

CHAPTER 20

Setup instructions (This is from Data Carpentry (http://www.datacarpentry.org/R-genomics/))

R and RStudio are separate downloads and installations. R is the underlying statistical computing environment, but using R alone is no fun. RStudio is a graphical integrated development environment (IDE) that makes using R much easier and more interactive. You need to install R before you install RStudio. After installing both programs, you will need to install the tidyverse package from within RStudio. Follow the instructions below for your operating system, and then follow the instructions to install tidyverse and RSQLite.

20.1 Windows

20.2 If you already have R and RStudio installed

Open RStudio, and click on "Help" > "Check for updates". If a new version is available, quit RStudio, and download the latest version for RStudio. To check which version of R you are using, start RStudio and the first thing that appears in the console indicates the version of R you are running. Alternatively, you can type sessionInfo(), which will also display which version of R you are running. Go on the CRAN website and check whether a more recent version is available. If so, please download and install it. You can check here for more information on how to remove old versions from your system if you wish to do so.

20.3 If you don't have R and RStudio installed

Download R from the CRAN website. Run the .exe file that was just downloaded Go to the RStudio download page Under Installers select RStudio x.yy.zzz - Windows XP/Vista/7/8 (where x, y, and z represent version numbers) Double click the file to install it Once it's installed, open RStudio to make sure it works and you don't get any error messages.

20.4 macOS

20.5 If you already have R and RStudio installed

Open RStudio, and click on "Help" > "Check for updates". If a new version is available, quit RStudio, and download the latest version for RStudio. To check the version of R you are using, start RStudio and the first thing that appears on the terminal indicates the version of R you are running. Alternatively, you can type sessionInfo(), which will also display which version of R you are running. Go on the CRAN website and check whether a more recent version is available. If so, please download and install it.

20.6 If you don't have R and RStudio installed

Download R from the CRAN website. Select the .pkg file for the latest R version Double click on the downloaded file to install R It is also a good idea to install XQuartz (needed by some packages) Go to the RStudio download page Under Installers select RStudio x.yy.zzz - Mac OS X 10.6+ (64-bit) (where x, y, and z represent version numbers) Double click the file to install RStudio Once it's installed, open RStudio to make sure it works and you don't get any error messages.

CHAPTER 21

Using DEseq and EdgeR to find differentially expressed genes

The first step is to merge all count data files we got from the Htseq. Use the Join two data sets side-by-side on Galaxy and select output from Control2 and Control3 samples. Use Column one to join the data sets. After this is complete, take the resulting file, and combine with Temperate1 output. Repeat this for the next two data sets.

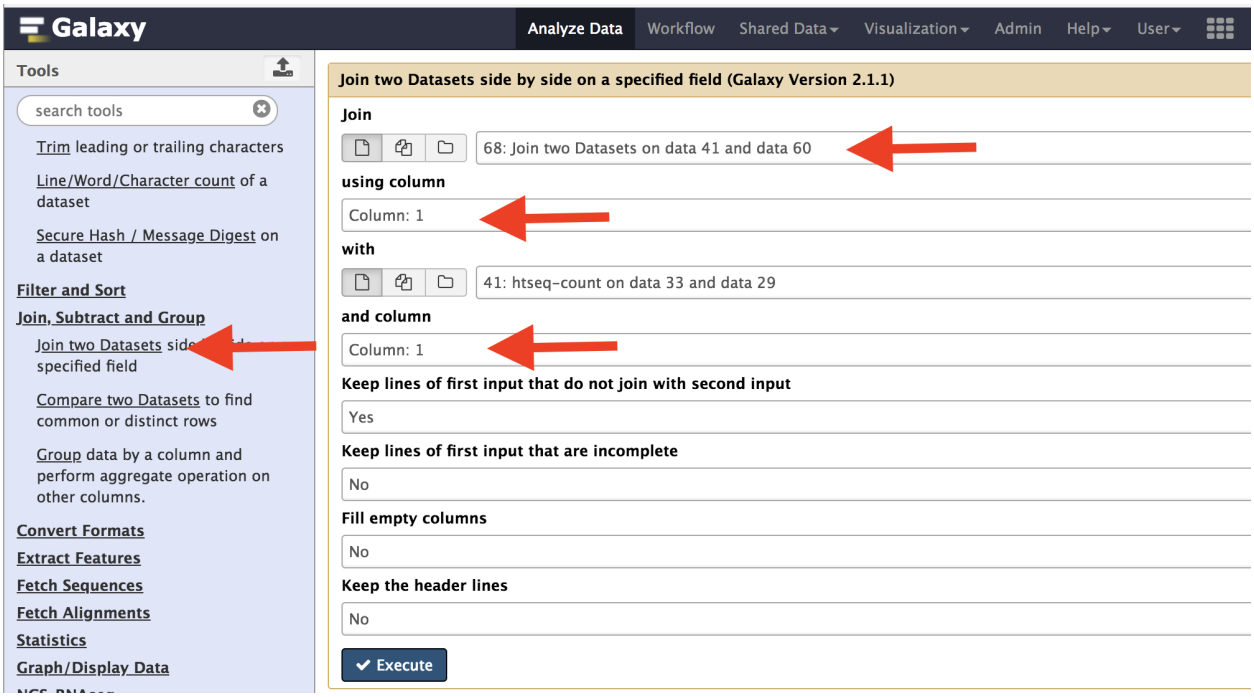
The resulting file would like this. Your actual number may be different, but should have 10 columns.

Download this file onto your computer and move it to folder called "Counts". Rename file "counts.tabular".

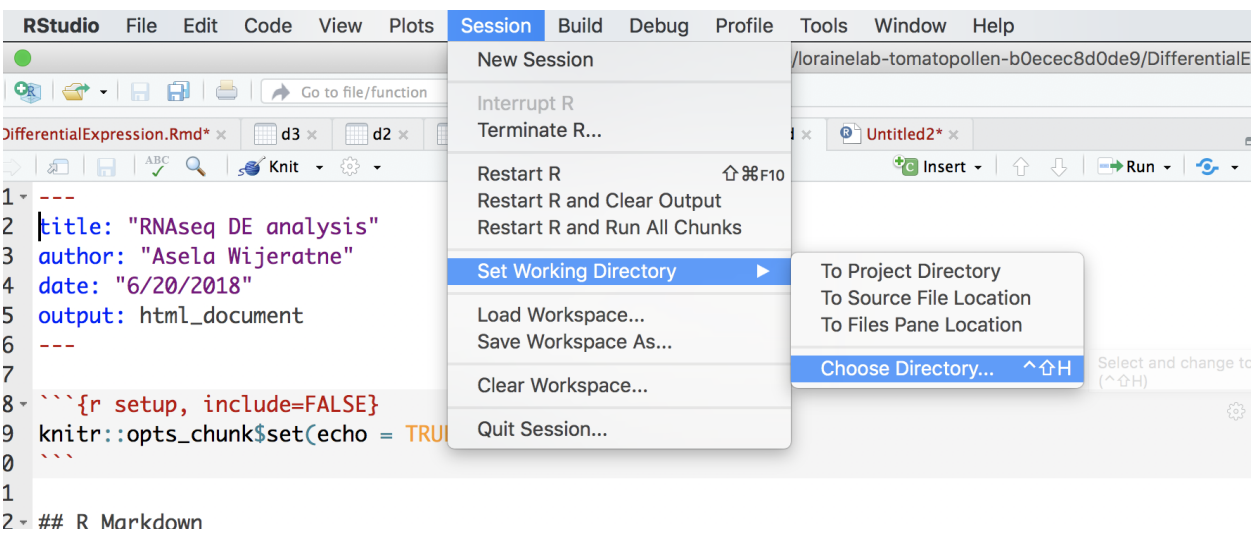
Open Rstudio and go to Session and select "Set Working Directory" and chose the folder that you just created.

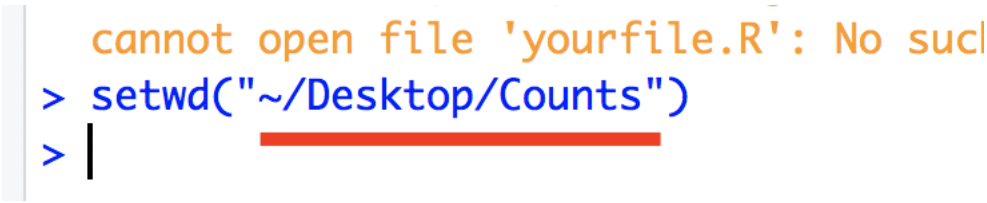
In the console, you will see the following message and the part underline in red is the path to your directory.

Replace your path in this portion of the following code "/Users/aselawijeratne/Desktop". Execute this to read file into R.



Geneid	TopHat on data 24	data 5	and data 4: accepted_hits								
mRNA:Solyc00g005000.3.1	0	mRNA:Solyc00g005000.3.1	0	mRNA:Solyc00g005000.3.1	0	mRNA:Solyc00g005000.3.1	0	mRNA:Solyc00g005000.3.1	0	mRNA:Solyc00g005000.3.1	0
mRNA:Solyc00g005005.1.1	0	mRNA:Solyc00g005005.1.1	0	mRNA:Solyc00g005005.1.1	0	mRNA:Solyc00g005005.1.1	0	mRNA:Solyc00g005005.1.1	0	mRNA:Solyc00g005005.1.1	0
mRNA:Solyc00g005040.3.1	0	mRNA:Solyc00g005040.3.1	0	mRNA:Solyc00g005040.3.1	0	mRNA:Solyc00g005040.3.1	0	mRNA:Solyc00g005040.3.1	0	mRNA:Solyc00g005040.3.1	0





```
d1=read.delim('/Users/aselawijeratne/Desktop/Counts/counts.tabular',␣
↪header = FALSE)
```

Format data using R.

```
#select only column with data
d1 <- d1[-c(3, 5, 7, 9)]

#Name the columns
colnames(d1) <- c("gene_names", "C2", "C3", "T1", "T2", "T3")

#get rid of the mRNA part infront of the gene name
row_names <- gsub("mRNA:", "", d1$gene_names)
#remove the last trailing ".1" from gene names
row_names <- gsub('.{2}$', '', row_names)

#assign row_name vector to the row names of the data.
row.names(d1) <- row_name

#remove unformated gene names.
d1$gene_names <- NULL
```

Import necessary libraries.

```
library(edgeR)
library(DESeq2)
```

Filter data.

```
#Filter data with rowsum < 10
d1$rowsum <- rowSums(d1)
#Low count filtered
d1_filterd <- d1[d1$rowsum > 10, ]
```

Create a group file and normalize data using EdgeR.

```
conds <- c(rep("C", 2), rep("T", 3))
y <- DGEList(counts=d1_filterd, group=conds, remove.zeros=TRUE) #␣
↪Constructs DGEList object

dge=calcNormFactors(y)
```

A multi-dimensional scaling (MDS) plot to see the similarity among samples.

```
# color for controls
cn.color='blue'
# color for treatments
```

(continues on next page)

(continued from previous page)

```
tr.color='brown'
# define a title for the plot
main='MDS Plot for Count Data'
#par(las=1) # makes y axis labels horizontal not vertical
colors=c(rep(cn.color,2),rep(tr.color,3))
plotMDS(dge,main=main,labels=colnames(dge$counts),
col=colors,las=1)
```

Hierarchical clustering can also be used to check how different samples are. As you can see, sample T2 is very different from the rest.

```
>normalized.counts=cpm(dge)
>transposed=t(normalized.counts) # transposes the counts matrix
>distance=dist(transposed) # calculates distance
>clusters=hclust(distance) # does hierarchical clustering
>plot(clusters) # plots the clusters as a dendrogram
```

|image1|

Differential expression analysis

```
y <- estimateCommonDisp(y) # Estimates common dispersion
y <- estimateTagwiseDisp(y) # Estimates tagwise dispersion
et <- exactTest(y, pair=c("C", "T")) # Computes exact test for the␣
↪negative binomial distribution.
topTags(et, n=4)
```

Convert data into a dataframe and use dfr and fold change to select genes.

```
edge <- as.data.frame(topTags(et, n=50000))
edge2fold <- edge[edge$logFC >= 1 | edge$logFC <= -1,]
edge2foldpadj <- edge2fold[edge2fold$FDR <= 0.01, ]
```

CHAPTER 22

DEseq analysis

Create matrix for DESeq2 and prepare data for DEseq

```
d1_matrix <- data.matrix(d1_filterd) #Create matrix for DESeq2

genotype <- data.frame(Genotype=conds) # create a dataframe for groups



dse <- DESeqDataSetFromMatrix(countData = d1_matrix,
                      colData = genotype,
                      design = ~ Genotype)

colData(dse)$Genotype <- factor(colData(dse)$Genotype,
                        levels=c("C","T"))
```

Differential expression analysis

```
dse <- DESeq(dse)
dse <- DESeq(dse)

ddsLocal <- estimateDispersions(dse, fitType="local", maxit =500)
ddsLocal <- nbinomWaldTest(ddsLocal)
res <- results(ddsLocal)
```

Order the data using p-adjusted value.

```
res <- res[order(res$padj),]
head(res)
res <- na.omit(res)
```

Use dfr and fold change to select genes.

```
deseq_2fold <- res[res$log2FoldChange >= 1 | res$log2FoldChange <= -1,]
deseq_2foldpadj <- data.frame(deseq_2fold[deseq_2fold$padj <= 0.01, ])
bothDF <- merge(deseq_2foldpadj, edge2foldpadj, by="row.names", all.
↪x=TRUE)
```

Writing results files. You can open "edgr_deseq2.txt" file in Excel if you want to look at it.

```
write.table(as.data.frame(bothDF), file="edgr_deseq2.txt", sep="\t",␣
↪quote = F, row.names=F, col.names=TRUE)
```

CHAPTER 23

Combine DESeq and EdgR to make Venn diagram

Count how many gene from each analysis and make a Venn diagram. You need to have overLapper.R file (down loaded from Bb) in Counts folder.

```
source("overLapper.R")

setlist <- list(edgeR=rownames(edge2foldpadj), DESeq=rownames(deseq_
↪2foldpadj))
OLlist <- overLapper(setlist=setlist, sep="_", type="vennsets")
counts <- sapply(OLlist$Venn_List, length)
vennPlot(counts=counts)
```

CHAPTER 24

GOseq analysis

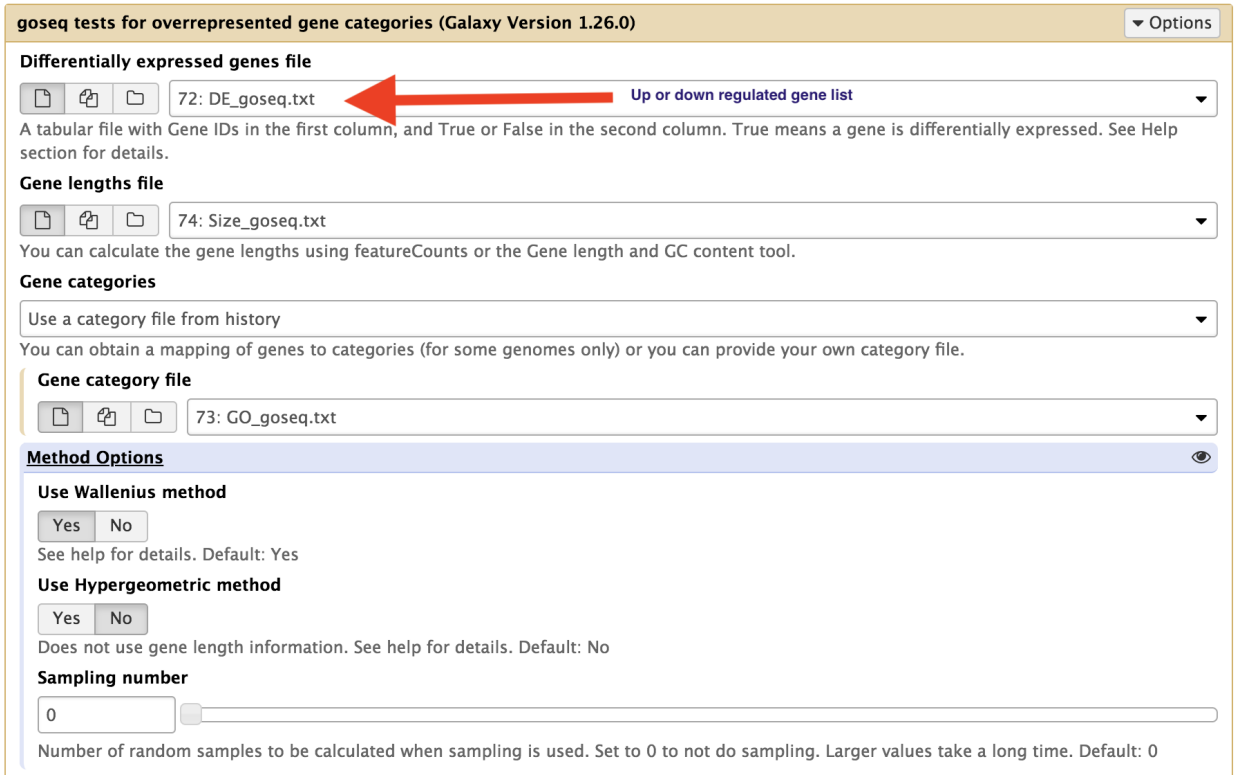
1. Ge the up and down regulated gene list. "bothDF" is the dataframe that contains both up and down-regulated genes from both EdgeR and DEseq2.

```
bothDF_down <- bothDF[bothDF$log2FoldChange <= -1,]
bothDF_up <- bothDF[bothDF$log2FoldChange >= 1,]
```

Convert to these dataframes into table with True or False values. Write the table to local directory.

```
DE_list_boolup <- as.data.frame(row.names(d) %in% bothDF_up$Row.names,
      row.names = row.names(d))
DE_list_booldown <- as.data.frame(row.names(d) %in% bothDF_down$Row.
↪names,
                        row.names = row.names(d))

write.table(DE_list_boolup,file="DE_goseq_up.txt",row.names=T,sep='\t',
↪quote=F,
    col.names = F)

write.table(DE_list_booldown,file="DE_goseq_down.txt",row.names=T,sep=
↪'\t',quote=F,
col.names = F)
```

2. Perform the GOseq analysis in Galaxy. You will need to perform the analysis for up and down regulated genes separately.



3. Combine gene descriptions with up and down regulated genes. You can get the S_lycopersicum_Feb_2014.bed file from the Dropbox link on Bb.

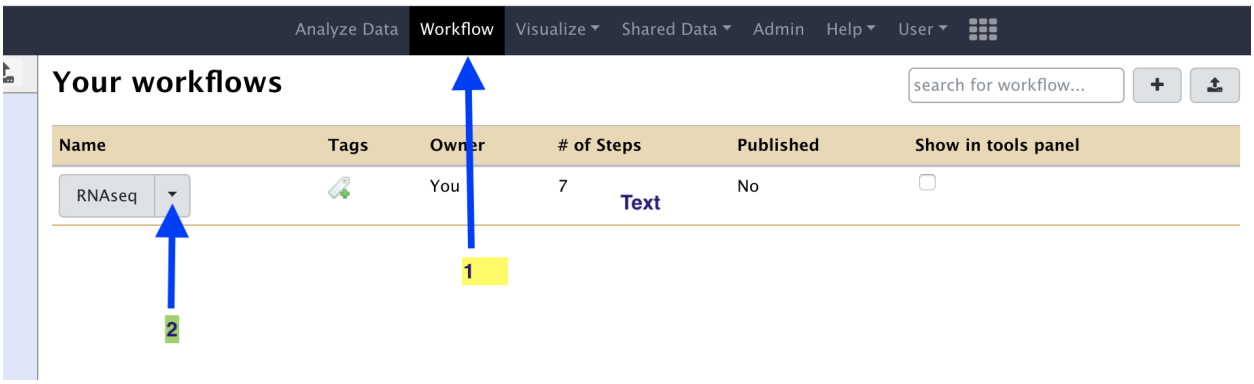
```
annots_file <- 'S_lycopersicum_Feb_2014.bed'
# keep gene id and gene description columns
annots <- read.delim(annots_file,sep='\t',header=F)[,13:14]
# name the columns
names(annots) <- c('gene','description')
# combine gene expression and annotations
bothDF_genedesc  <- merge(bothDF,annots, by.x = "Row.names",by.y='gene
↪')
```

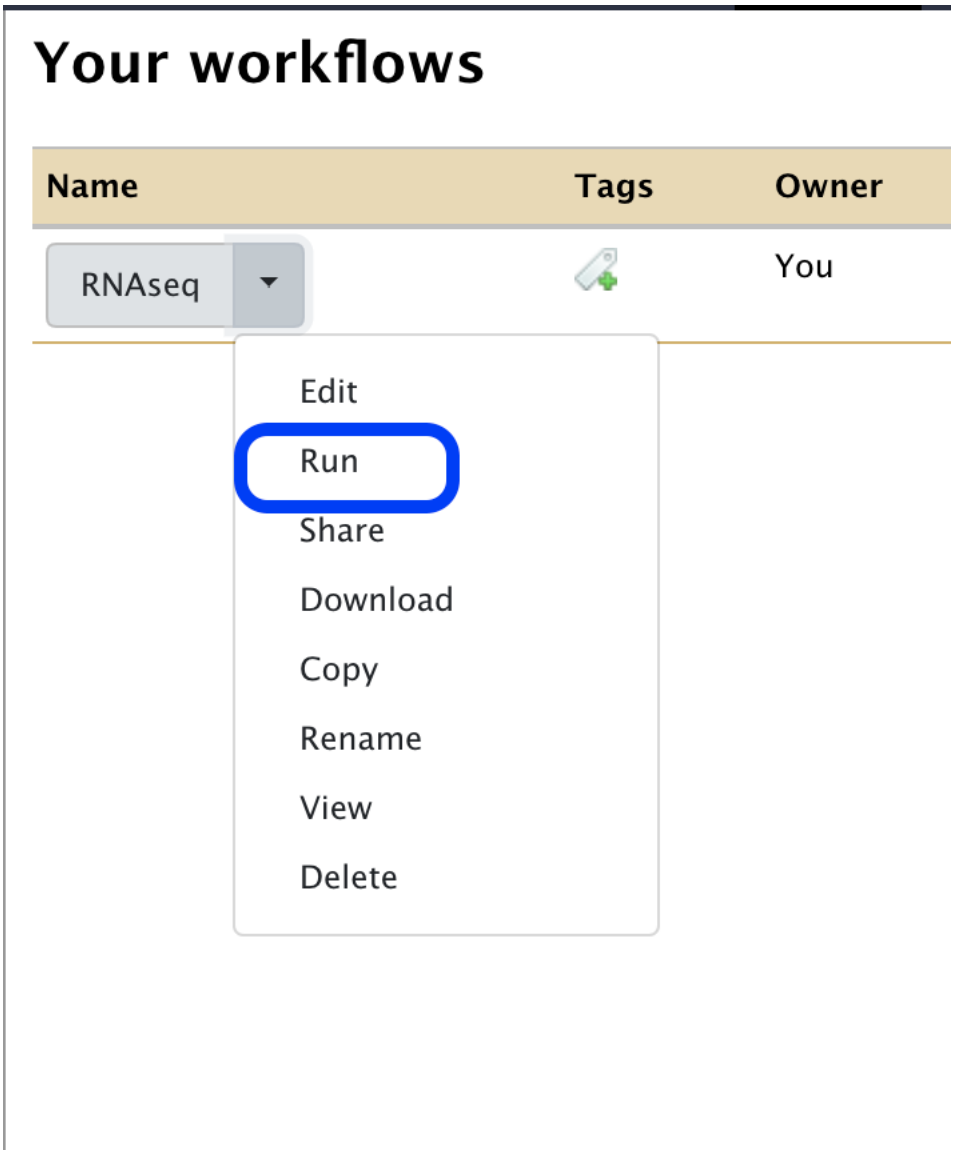
4. To order your data using FDR, you use the following command in R.

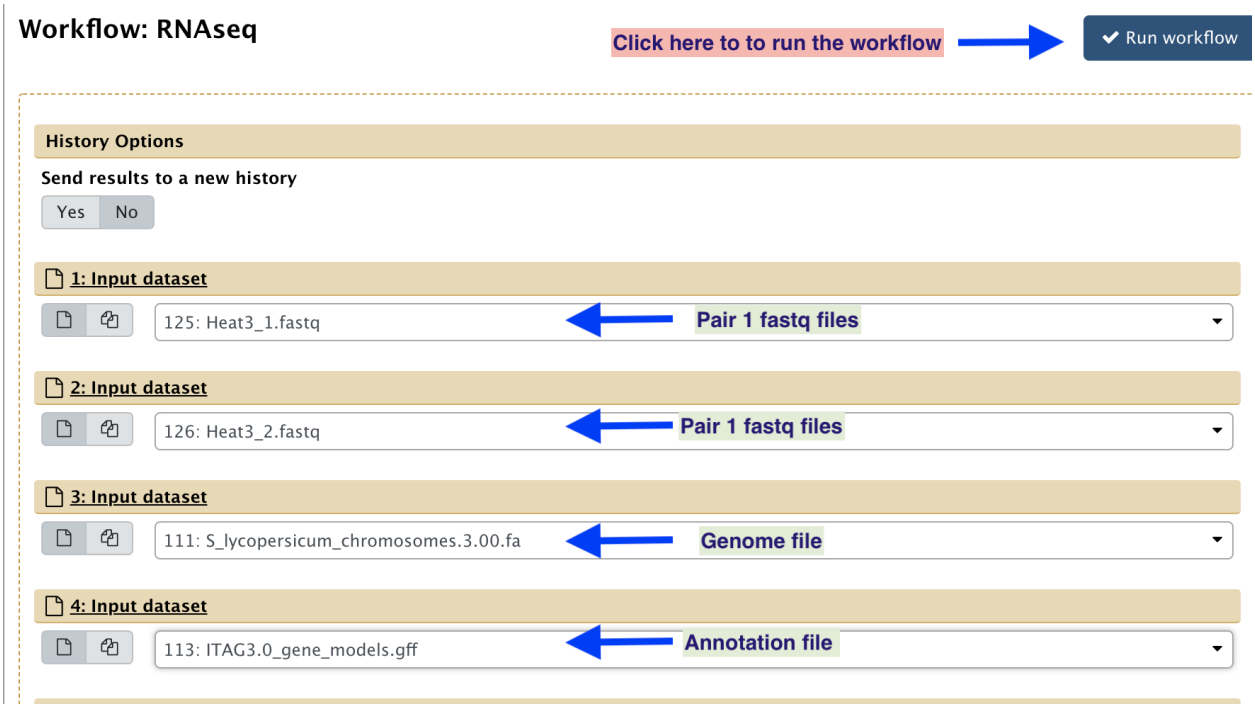
```
bothDF_genedesc  <-  bothDF_genedesc[order(bothDF_genedesc$FDR), ]
```

CHAPTER 25

Run RNAseq analysis as a workflow




Your workflows
Name
Tags
Owner
RNAseq
You
Edit
Run
Share
Download
Copy
Rename
View
Delete


Workflow: RNAseq
Click here to to run the workflow
Run workflow
History Options
Send results to a new history
Yes
No
1: Input dataset
125: Heat3_1.fastq
Pair 1 fastq files
2: Input dataset
126: Heat3_2.fastq
Pair 1 fastq files
3: Input dataset
111: S_lycopersicum_chromosomes.3.00.fa
Genome file
4: Input dataset
113: ITAG3.0_gene_models.gff
Annotation file

CHAPTER 26

Indices and tables

- genindex
- modindex
- search