Quality Control of Illumina Data using Galaxy

August 18, 2014

Contents

1 Introduction

1.1 What is Galaxy?

Galaxy is a web-based interface that allows users to create complex computational pipelines to analyze biological data. Galaxy is designed to help you create reproducible workflows that can be used with multiple datasets, shared with others and published. Common bioinformatics software such as BLAST, BWA and GATK can be accessed though the Galaxy interface along with many other tools for converting between different formats, manipulating data and basic statistics.

1.2 Galaxy at MSI

There are many instances of Galaxy, the one available to you though MSI is maintained by MSI and connects directly to the computational resources at MSI. The tools available will vary depending on which instance of Galaxy you use. While transferring workflows from one instance of Galaxy to another is easy, MSI has no control over which tools are available in other Galaxy instances. If there is a tool that you have used in a different instance of Galaxy that is not available in the MSI instance send a request to help@msi.umn.edu.

1.3 Scope of this tutorial

- Give participants experience with the basic functionality of Galaxy
  - Accessing Galaxy at MSI
  - Galaxy layout
  - Loading files into current Galaxy history
  - Creating a workflow
  - Sharing histories and workflows
  - Where to get more information
- Basic processing and quality control on Illumina sequencing data
  - Evaluating read quality
  - Adapter removal
  - Low quality read removal
  - Read trimming

1.4 Where to get more information

- From other Galaxy users: `https://wiki.galaxyproject.org/`
- From MSI: `https://www.msi.umn.edu/content/bioinformatics-analysis`

2 Getting Started With Galaxy

Sections of Galaxy

Galaxy has three main sections; Tools Pane, Histories Pane, and the Center Pane.

- **Tools Pane**
  - Found on the left side of the browser.
  - Contains all of the different tools that can be used within Galaxy. These include tools that do simple text manipulations and arithmetic to tools with more complex functions specific to the analysis of next generation sequence data and statistics.
  - Combining these different tools allows you to analyze your data.
  - Tools are organized into several heading or they can be found using the search bar at the top of the pane.
- **Histories Pane**
  - Found on the right side of the browser.
  - Contains the history of the tools you have used and the results.
  - Histories can be saved, shared and turned into workflows that can also be saved, shared and reused.
- **Center Pane**
  - Found in the center of the browser.
  - When using a tool the options for that tool and information about the tool will be in the Center Pane.
  - Clicking on the *Eye Icon* in the history pane will give of view of the data in the Center Pane.

2.1 Accessing Galaxy at MSI

a) Open a web browser and navigate to the MSI Galaxy `galaxy.msi.umn.edu`

b) Log in with your MSI username and password

c) Tools Pane

d) Center pane

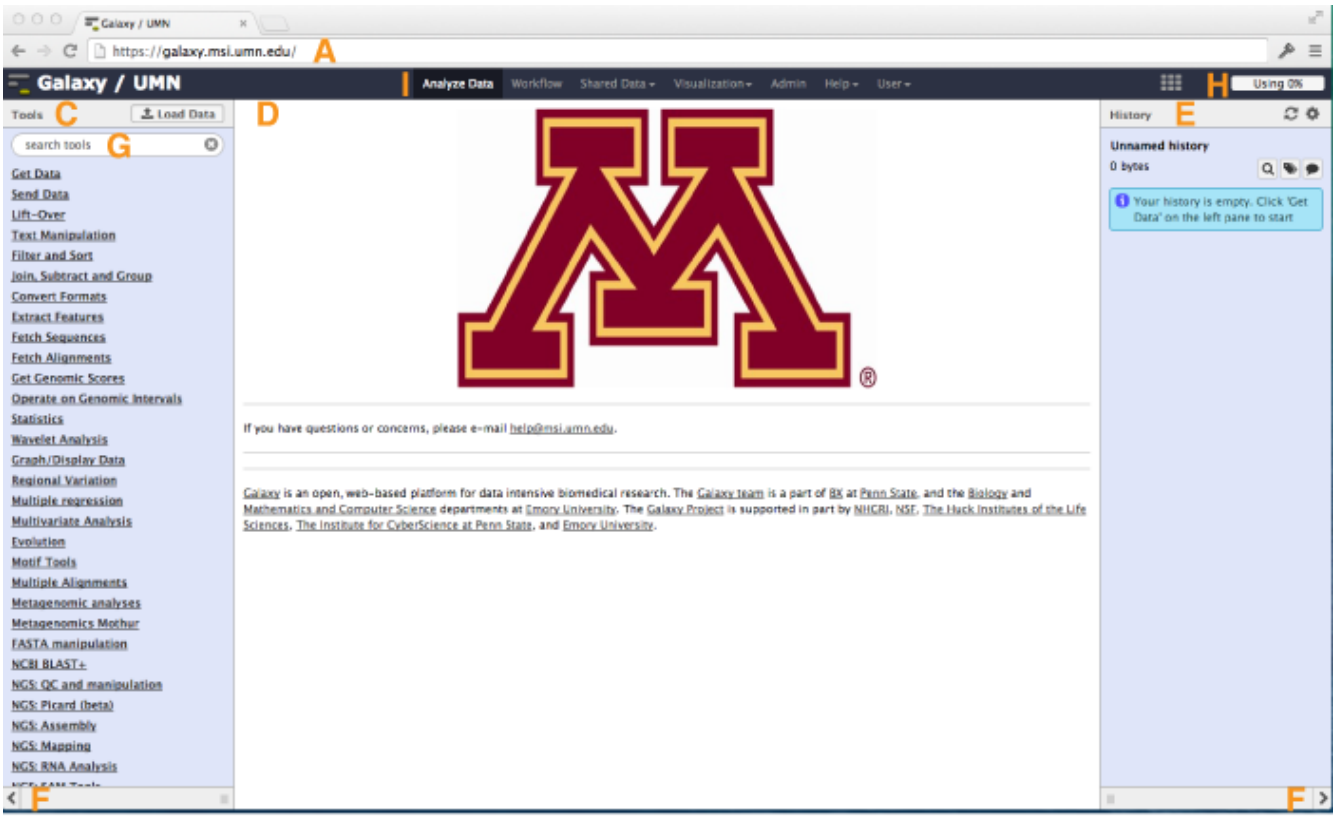
e) History Pane

f) The side panels can be collapsed via arrows in the bottom corner to provide a better view of the Center Panel.

g) Search bar to find tools.

h) The total quantity of data your group has stored in Galaxy is displayed in the top right corner.

i) You can always get back to the main screen using *Analyze Data* in the top menu bar.



2.2 Import FASTQ Files From Data Library

Getting data into MSI Galaxy- Data Libraries

- Sequencing Data from UMGC
  Sequencing data from UMGC can be accessed in Galaxy though the creation of a data library. In general, each PI with access to Galaxy account will have one data library that can contain many different pieces of data. When you or your PI receives an email from UMGC indicating that your sequencing data is available you can have that data moved into your PI's Galaxy data library by forwarding the email to help@msi.umn.edu with a request to add the data to Galaxy. You can then access the sequencing data library from the *Shared Data* tab in the blue bar at the top of the Galaxy page. If your PI doesn't currently have a data library a new one will be created the first time you request to have data added to Galaxy.

- External Data
  The *Get Data* heading in the Tool Pane is a good resource for obtaining external data from public databases such as the UCSC genome browser and SRA. You can also upload small (<2GB) files directly from your computer. When data is uploaded using the tools under *Get Data* they will appear in your current history.

- Larger External Datasets
  Data files that are larger than 2GB will have to be placed into a data library to be accessed in Galaxy. In your groups home directory there is a galaxy folder (/home/yourGroup/galaxy). To get data into your PI's data library move it into the galaxy folder in your groups home directory then send a ticket to help@msi.umn.edu with the location of the data to be added to your PI's data library.

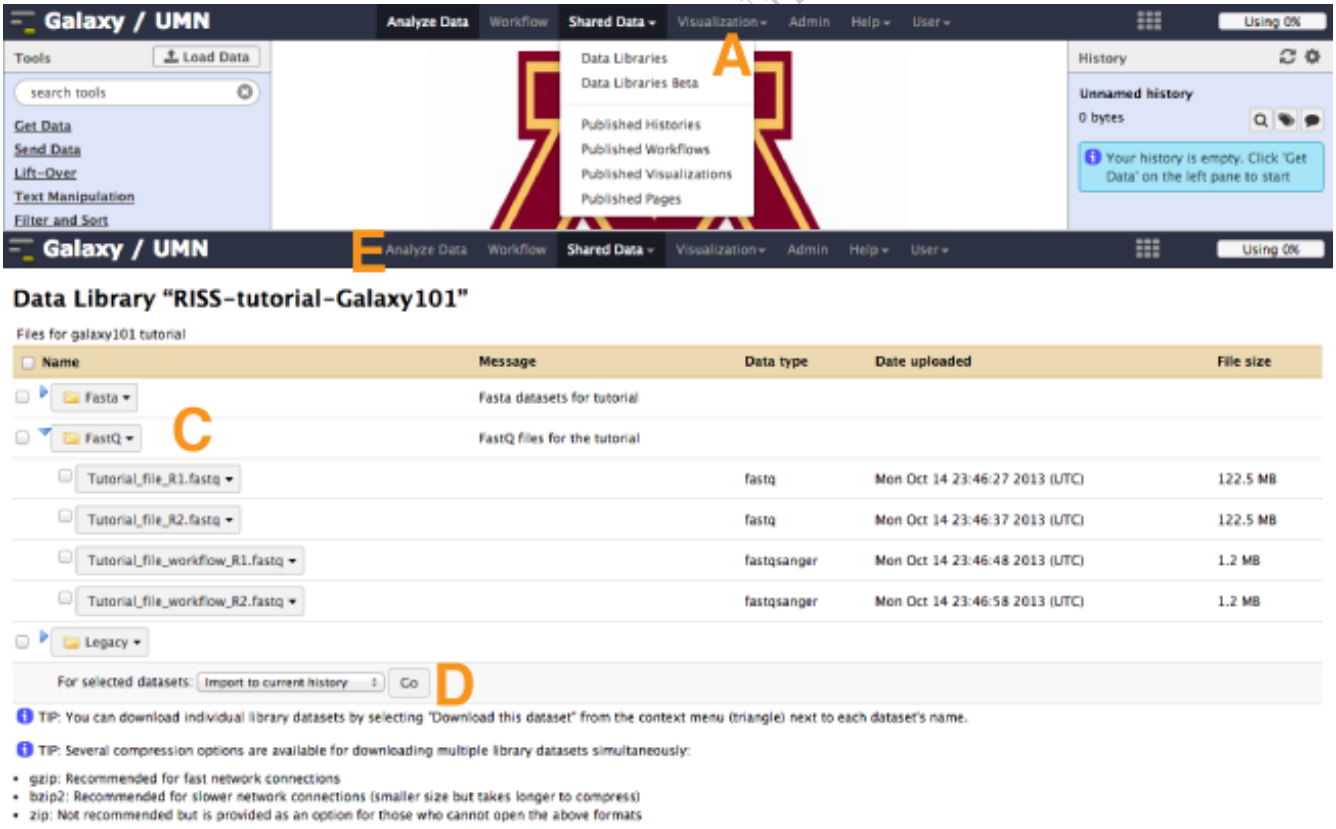
a) At the top of the screen select *Shared Data* then in the menu *Data Libraries*

b) Select *RISS-tutorial-galaxy101* from the list of data libraries

c) Expand the *FastQ* folder and check the boxes next to the first two files, Tutorial_file_R1.fastq and Tutorial_file_r2.fastq

d) Select *Go* next to *Import to current history* below the data files to move the data to your current history.

e) Select *Analyze Data* in the blue bar to move back to the main Galaxy view.



2.3 Set File Attributes

Attributes

Setting the file attributes will tell the different tools in Galaxy what format the data is in. Galaxy does some work to auto detect the files that can be used as inputs for different tools. If you find that the file you want to use as an input is not available in a drop down menu check to see if you have set the file attributes. Information about different files types can be found though the **USCS genome browser** and from **Current Protocols in Bioinformatics**.

Special note about FASTQ

FASTQ files contain quality information for each sequenced base encoded using the characters found in the fourth line of each block. The preferred encoding for MSI Galaxy is Sanger. If you are looking at Illumina data created in 2012 or later your FASTQ files already using the Sanger encoding. If you sequencing was done before 2012 then you should use FASTQ Groomer to convert to the Sanger encoding (Sanger & Illumina 1.8+).

Canonical genomes

Both Mouse (mm9) and Human (hg19, hg18) have canonical versions in Galaxy. For most NGS analyses you will want to use the canonical versions of the genome if available. These genomes contain only the standard chromosomes (i.e., somatic, sex and mitochondria) and do not include parts of the genome that have unknown locations, haplotype specific chromosomes or random chromosomes.

a) In the History Pane click on the *pencil icon* next to Tutorial_file_R2.fastq. This will bring up the files Attributes in the Center Pane.

b) This is a human dataset so select *Human hg19 in GATK canonical* in the drop down menu under *Database/Build:*. You can scroll or if you begin to type "hg19" then you will only see the options with "hg19" in the name.

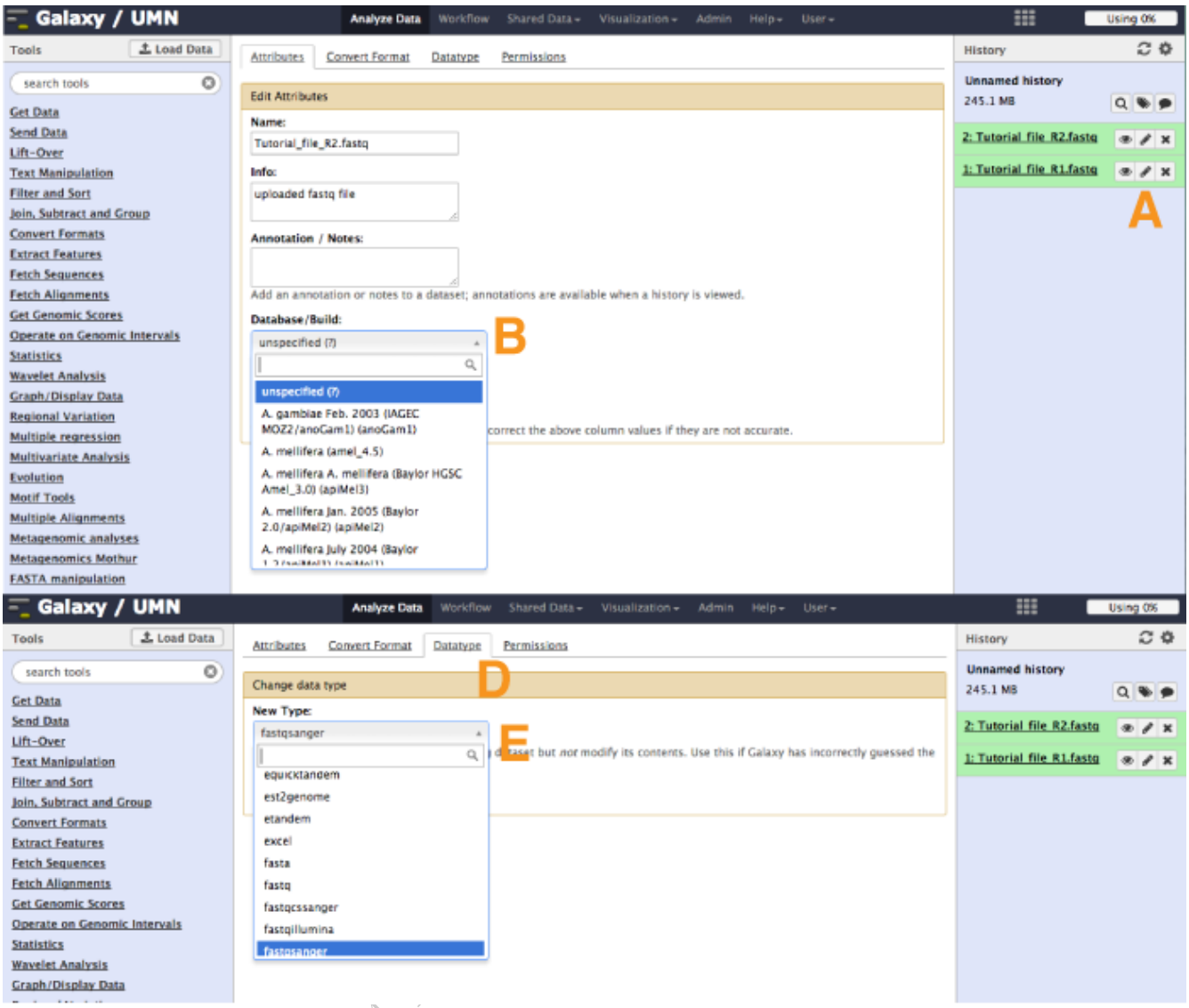
c) Click *Save*

d) Switch to the *Datatype* tab by selecting it from the top of the Center Pane

e) Select *fastqsanger* from the drop down menu. You can scroll or if you begin to type "faster" then you will see the options with "fastq" in the name. **NOTE:** do not select fast**qc**sanger.

f) Click *Save*

g) Change the attributes and data type for the other fastq file to match.



3 Evaluating FASTQ File Quality

FASTQ Format and Quality Scores

This tutorial is geared towards Illumina data in FASTQ format, other sequencing methods (i.e., Roche 454) may produce reads with a different patterns of errors or a different file format. Quality control tools for other NGS data types can be found under the *NGS: QC and manipulation* heading in the Tool Pane.

A sequence record in a FASTQ file consist of four lines 1) an @accession line 2) sequence data 3) + place holder line 4) quality score line. FASTQ quality scores encode the estimated chance of a miscalled base at each location. Single ASCII characters are used to encode the quality scores, as opposed to raw numbers, so that there is always a 1-to-1 relationship between the number of bases in the read and the length of the quality score. Quality score reflect the probability that a base call was incorrect, calculated as a Phred quality score (Phred Q = -10log(p), where p is the probability that the inferred base is incorrect). The higher the Phred score the smaller the

probability that the base call was incorrect. A Phred score of 10 indicates a 1 in 10 chance of an incorrect base call while a oared score of 50 indicates a 1 in 100,000 chance of an incorrect base call.

Unfortunately, FASTQ files from different sources sometimes encode quality scores slightly differently. Sanger and current Illumina FASTQ format uses a Phred+33 encoding, which means that the lowest Phred score of 0 is encoded as ASCII character 33 (!), while Solexa and pre-2012 Illumina software uses Phred+64 encoding (Phred 0 encoded as @). But in all cases, the higher the Phred quality scores the higher quality the base call. In Galaxy you can use FASTQ Groomer to ensure your data is in the Sanger/Illumina 1.8 + encoding.

FastQC Metrics

- Basic Statistics
  Gives the name of the input file, encoding used for the quality score, total sequence count, average sequence length and GC content (%).
- Per Base Sequence Quality
  A important figure showing the average quality score at each position across all reads. In general, quality scores are lower at the start and each of reads. Sudden dips in the middle of a read can signify failed cycles in the sequencing run (machine errors).
- Per Sequence Quality Scores
  Histogram charting the average quality across a read. Low quality reads can be removed but a majority (at least 75%) of you data should be of high quality.
- Per Base Sequence Content
  The frequency of each nucleotide at each position across all of the reads. Extremely high nucleotide bias can be a sign of trouble, short stretches with high bias can be caused by the presence of linkers, barcodes or adapter contamination. There is usually some minor bias in the first 11-13bp of RNA-seq data due to not-quite random hexamer sequence priming but this bias is accounted for in the downstream analysis.
- Per Base GC Content
  Average GC content (%) at each position along the reads. GC content should be stable across the read and large changes indicate issues.
- Per Sequence GC Content
  This figure will show you both the theoretical distribution of GC content and the GC content of your data. These distributions should be similar.
- Per Base N Content
  Rate of ambitious base calls (N) for each position along the reads. This count should be very low (<10), to many N calls indicates issues with the sequencing run (usually machine errors).
- Sequence Length Distributions
  Histogram of the sequence lengths. Illumina reads that have not been trimmed will all have the same length, once trimmed you want a majority of your reads to be full length and a small percentage to be shorter.

- Sequence Duplication Levels
  Frequency of exact sequence duplicates in the dataset. High duplication rates can be caused by PCR artifacts and/or low library diversity. Low levels of duplication can be removed but, high levels indicate issues with the library preparation.

- Overrepresented Sequences
  A list of overrepresented sequences if they exist in the data. These are the sequences that are contributing to the data in the Sequence Duplication Level graph.

- K-mer Content
  Shows the amount (% of reads) and sequence of overrepresented K-mers. High levels of overrepresented sequences usually arise from adapter contamination and these levels should drop after adapters are removed from your data.

3.1 Running FastQC

a) From the Tools Pane select the *NGS: QC and manipulation* header.

b) Select the *FastQC: Read QC* tool.

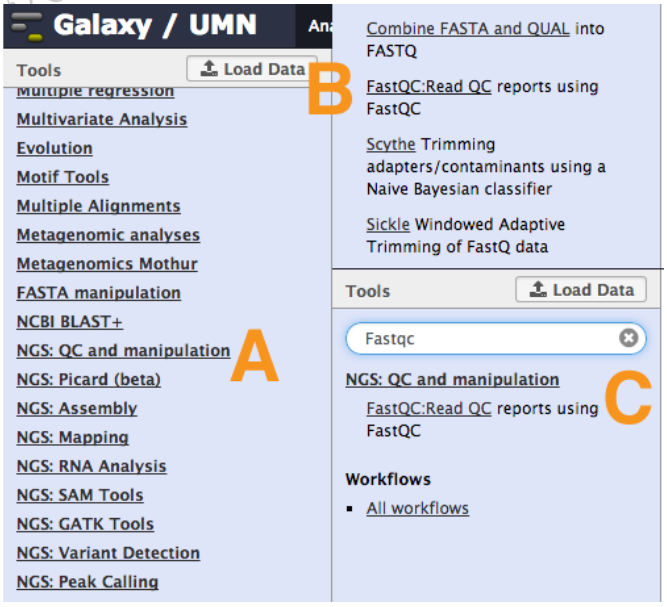
c) Alternatively, use the *search bar* at the top of the tool pane to find FastQC.

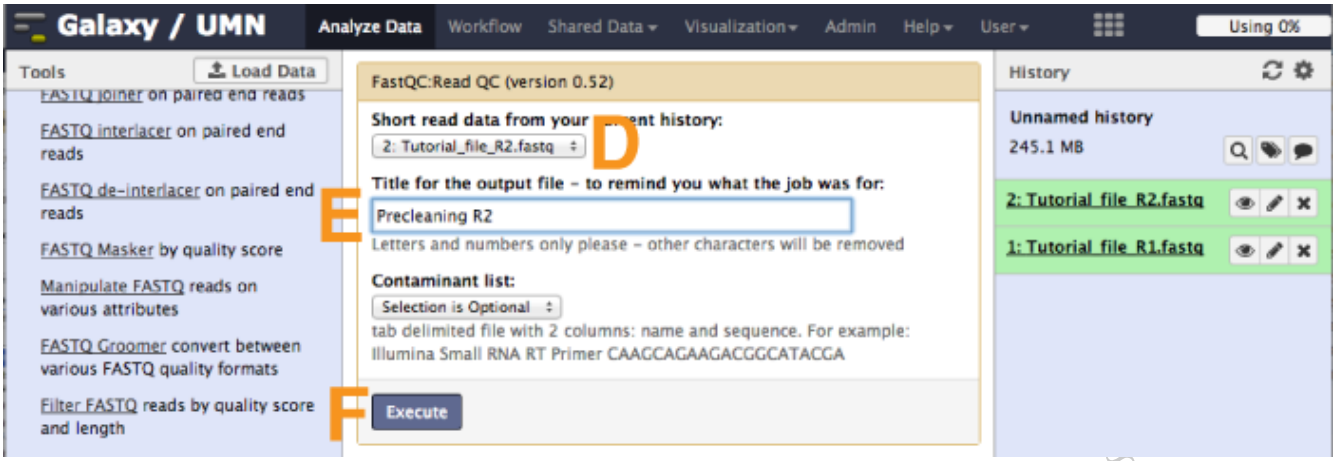
d) Select the file to analyze from the drop-down menu: Tutotial_file_R2.fastq

e) Rename the output file to something meaningful such as "Precleaning R2".

f) Select *Execute*

g) Repeat this process for the Tutorial_file_R1.fastq file changing the name of the output to reflect the use of the new file ("Precleaning R1").

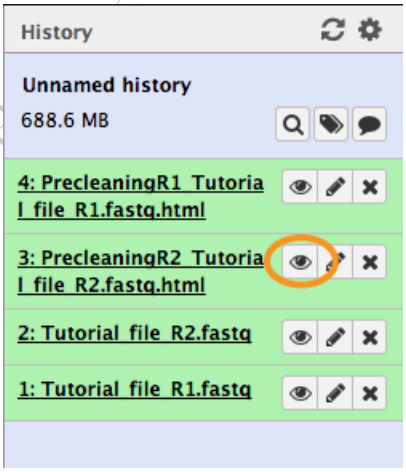




3.2 Viewing and Understanding FastQC results

In the following review of the FastQC results we will present the result of both the Left (_R1) and Right (_R2) FASTQ files side by side. We do this to accentuate the differences between the two datasets and to highlight the importance of checking the quality of both sets of reads. It is normal for one set of reads to be considerably different in quality from the other. Usually, the Left reads are higher quality due to being sequenced first.

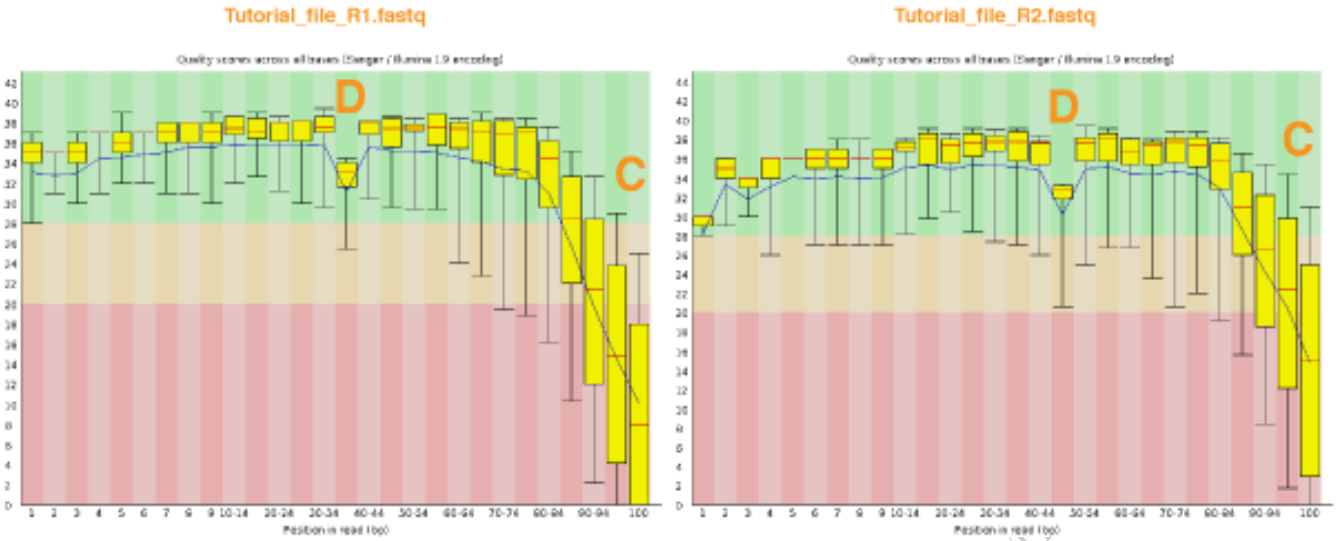
a) In the History Pane select the *Eye Icon* next to the name of the output from using the *FastQC: ReadQC* tool. This will allow you to view the results in the Center Pane.



b) Scroll to the "Per base sequence quality".

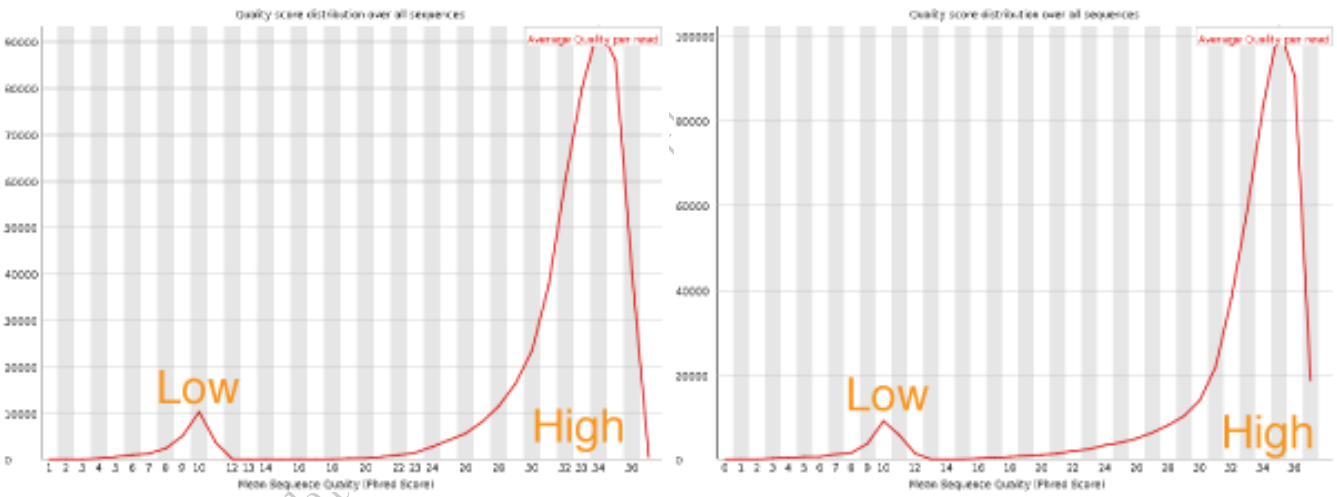
c) Note how the quality of the reads drops towards the 3' ends of the reads

d) Dips like this are indicative of a failed cycle on the sequencing machine.



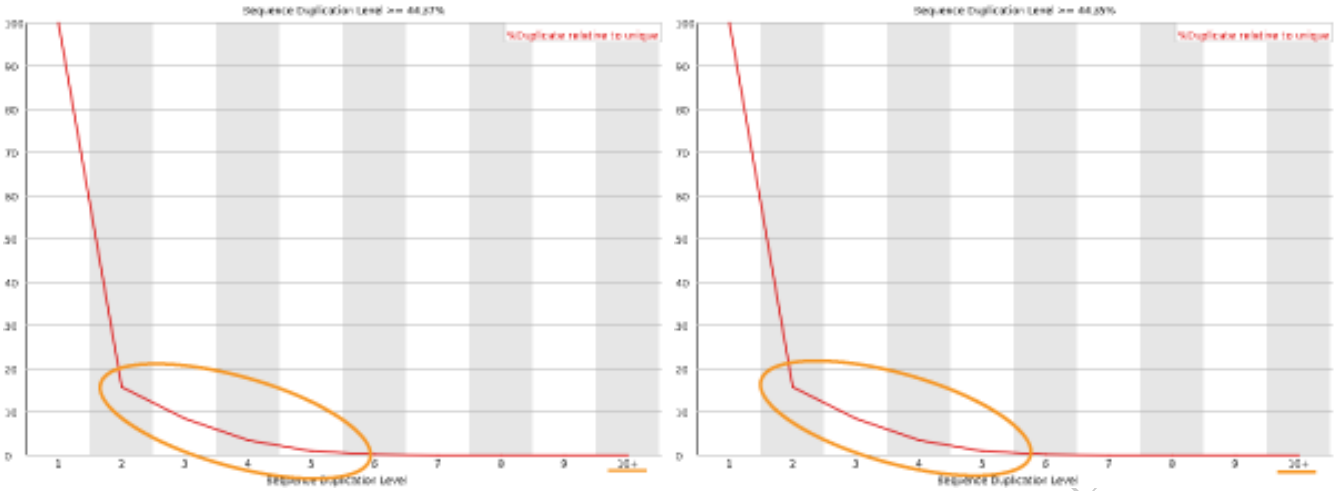
e) Scroll to "Per sequence quality scores".

f) Note the bimodal distribution with a population of low quality reads.



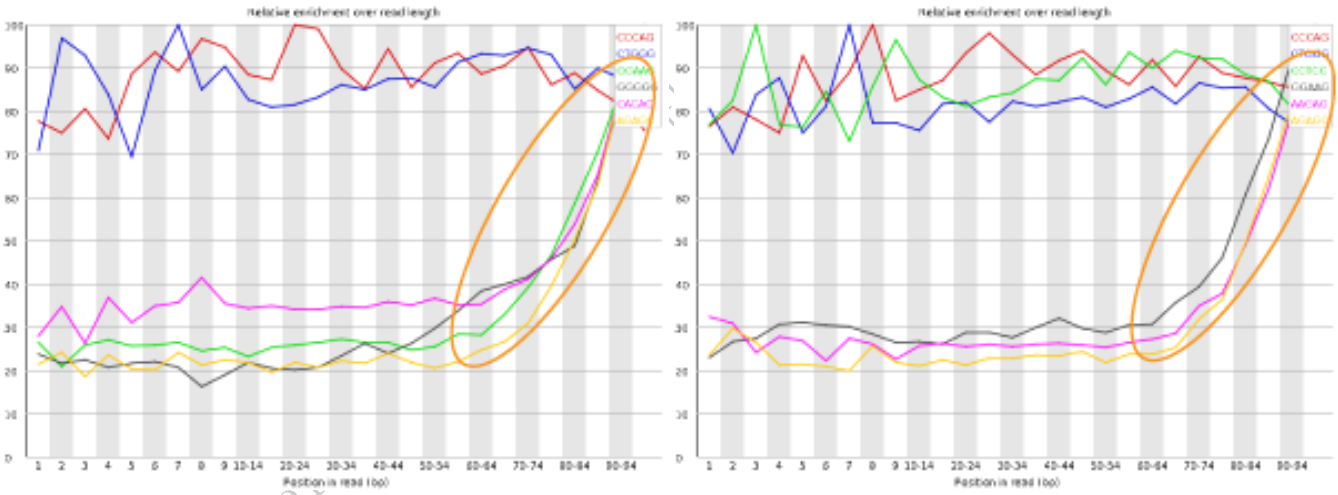
g) Scroll to "Sequence duplication levels".

h) Note the presence of duplicated reads, here up to 5 copies. Some duplication is expected and this is a relatively low level of duplication. Large number of sequences in the 10+ column would indicate issues.



i) Scroll down to “Kmer content”

j) Note the presence of over enriched k-mere at the 3’ end of the reads, this is indicative of 3’ adapter contamination. Adapter sequences may also be identified as an ”Overrepresented Sequence”.



4 Cleaning FASTQ Datasets

4.1 Why Is Cleaning Required?

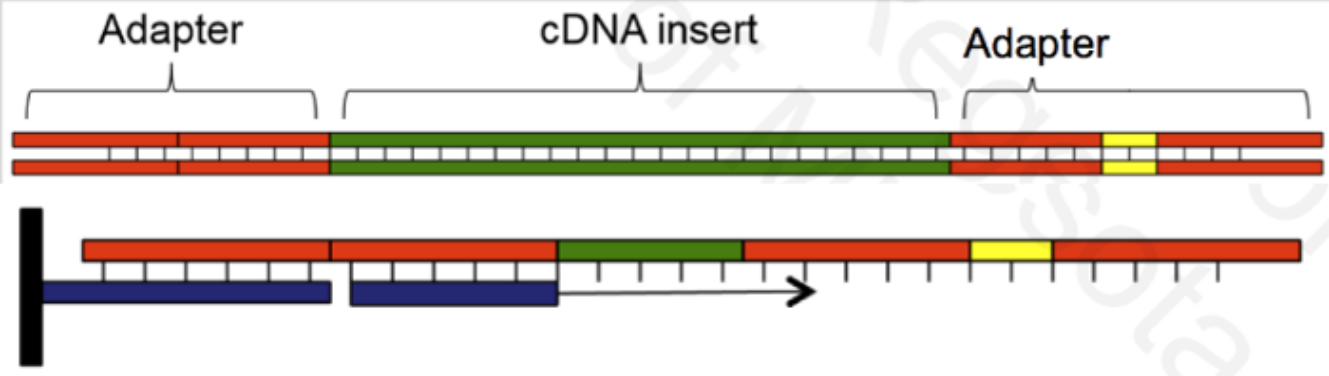
Low Quality Tails and Failed Cycles

For a variety of reasons, including decay of reagents as they sit on the sequencing machine, the quality of base calls tends to decrease as sequencing progresses. As a result the 5’ ends will tend to have higher quality than the 3’ ends and forward reads will tend to have better quality than reverse reads. Low quality base calls can impair the accuracy of mapping algorithms so it is important to to remove them. Low quality tails can be removed though the removal of the 3’ ends from all of the reads but, that would result in the removal of many reads that were of higher quality as well.

More sophisticated methods only remove the tails that show evidence of low quality. In Galaxy there are tools that can accomplish either style of read trimming.
Cycles fail because of sequencing machine error, such as failure to incorporate a base, or failure to image a specific region. For most analysis failed cycles can be ignored as they will not have large effects.

Adapter Contamination

Illumina libraries consist of the DNA of interest (green) with ligated adapter (red + yellow) on the 5' and 3' ends to proved priming site for the sequencing reactions. The forward adapter (left) provides a region that binds to the Illumina flow cell plate (blue) and a region to which the sequencing primer binds to start the sequencing reactions. The reverse adapter (right) has the same structure with the addition of a barcode sequence (yellow). Adapter contamination occurs when the DNA fragment of interest is shorter than the length of the sequencing read. This results in the opposite primer included in the sequence of the final read, leading to adapter contamination.
Removal of adapter contamination in Galaxy can be accomplished using *CutAdapt* and the sequence of the primers.



Illumina TrueSeq adapters:
Forward: 5' AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT 3'
Reverse: 5' GATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG 3'

4.2 Remove Low Quality Tails

a) Open the *NGS: QC and manipulation* section of the Tool Pane, or use the *search bar*.

b) Select *FASTQ Quality Trimmer*

c) In the *FASTQ drop down menu* select the Right FASTQ file, Tutorial_file_R2.fastq

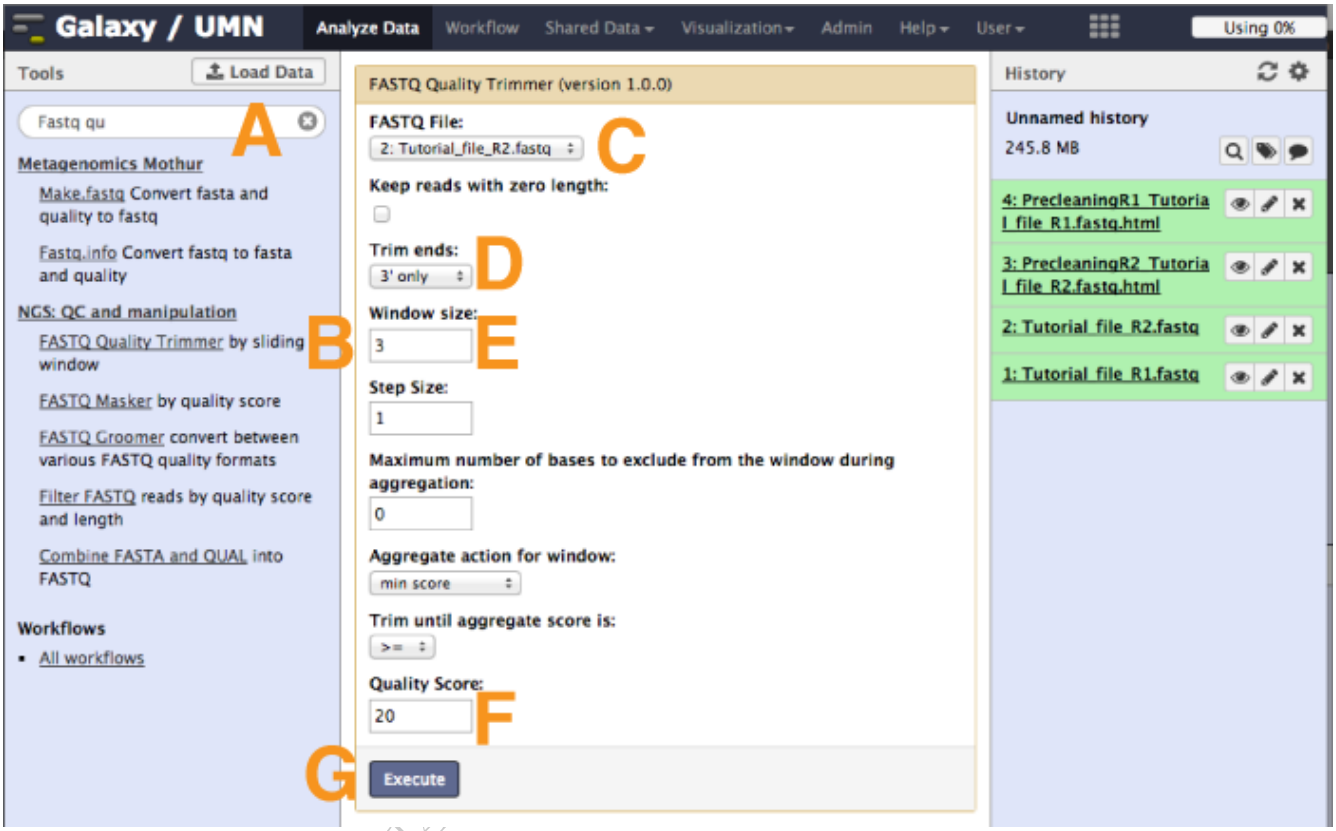
d) Set *Trim ends* to 3'only.

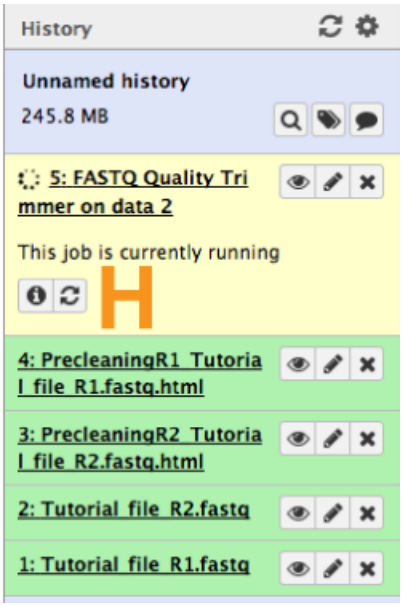
e) Set *Window size* to 3.

f) Set *Quality Score* to 20.

g) Select *Execute*.

h) Repeat these steps on the Left FASTQ file. To rerun a tool from your history you can select the underlined name of the output in the History Pane, expanding the box. Then select the *Recycle Icon* which will load the tool with the setting preloaded. For *FASTQ Quality Trimmer* instance you will just need to change the name of the input files.





4.3 Remove Adapter Contamination

a) Select *FASTA manipulation* from the Tool Pane.

b) Select *Cutadapt* to bring up the tool options in the Center Pane.

c) Under *Fastq file to trim:* select Right (_R2) FASTQ file that has already been trimmed for quality.

d) Select *Add new 3' Adapters*

e) Under *Choose 3' Adapters* select "TrueSeq Universal Adapter Reverse Complement"

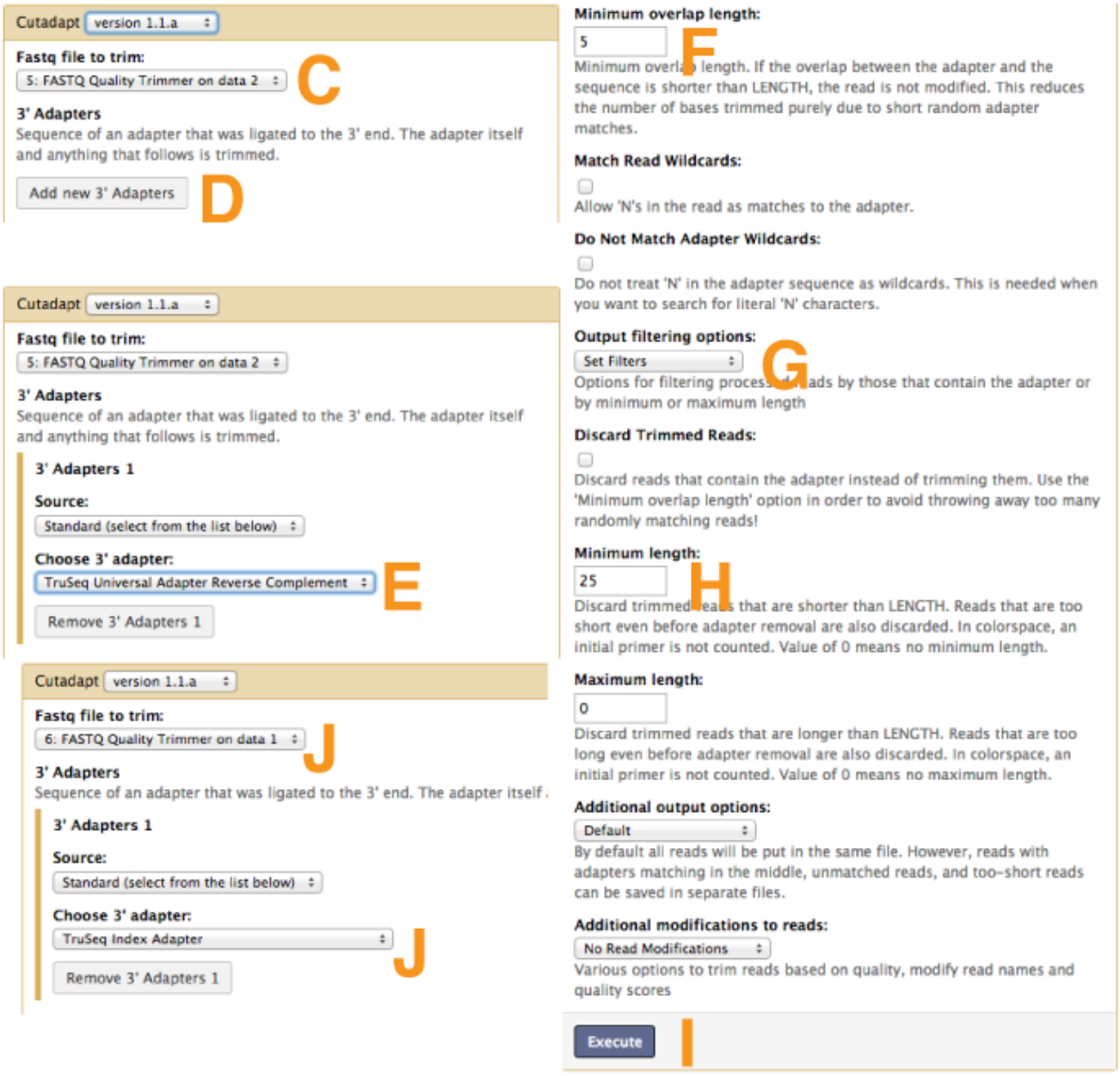
f) Set "Minimum overlap length" to 5

g) Set "Output filtering options:" to ''Set Filters".

h) Set "Minimum length" to 25

i) Select *Execute*.

j) Run *CutAdapt* again (*Recycle Icon*) using the Left (_R1) quality trimmed FASTQ file. Use "TrueSeq Index Adapter" as the 3' adapter.



k) *Cutadapt* will produce a report as well as the cleaned up FASTQ file.

l) Open the *Cutadapt* report by selecting the *Eye Icon* in the history pane.

m) Note the length distribution of removed sequence.

4.4 Resyncing Left and Right FASTQ Files

Trimming and other quality control measures can result in the removal of reads from the dataset. Additionally, some processing steps may change the order of the reads in the dataset. Many programs expect to find the same read names in the same order for both the Left and Right

FASTQ files. To ensure that the read names are in 'sync' MSI Galaxy has the *resync: Paired-end resynchronization* tool.

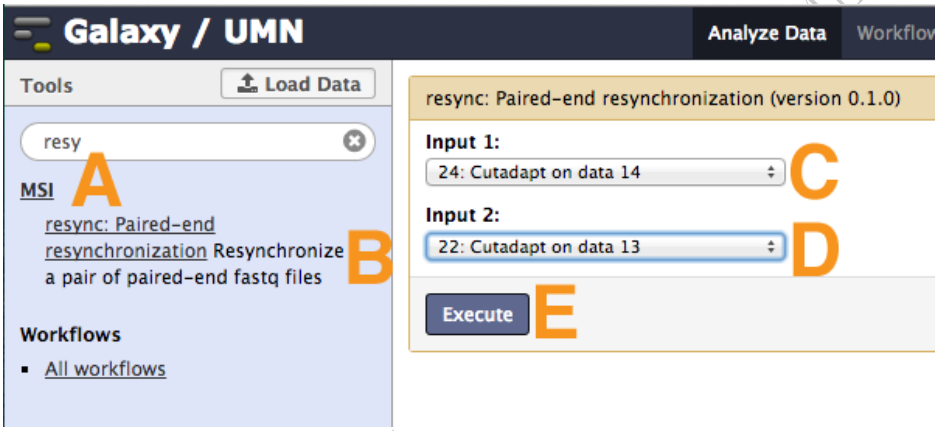
a) Select the *MSI* header from the Tool Pane.

b) Select *resync: Paired-end resynchronization*.

c) Select as *Input 1:* the quality and adapter trimmed Right FASTQ file

d) Select as *Input 2:* the quality and adapter trimmed Left FASTQ file

e) Select *Execute*

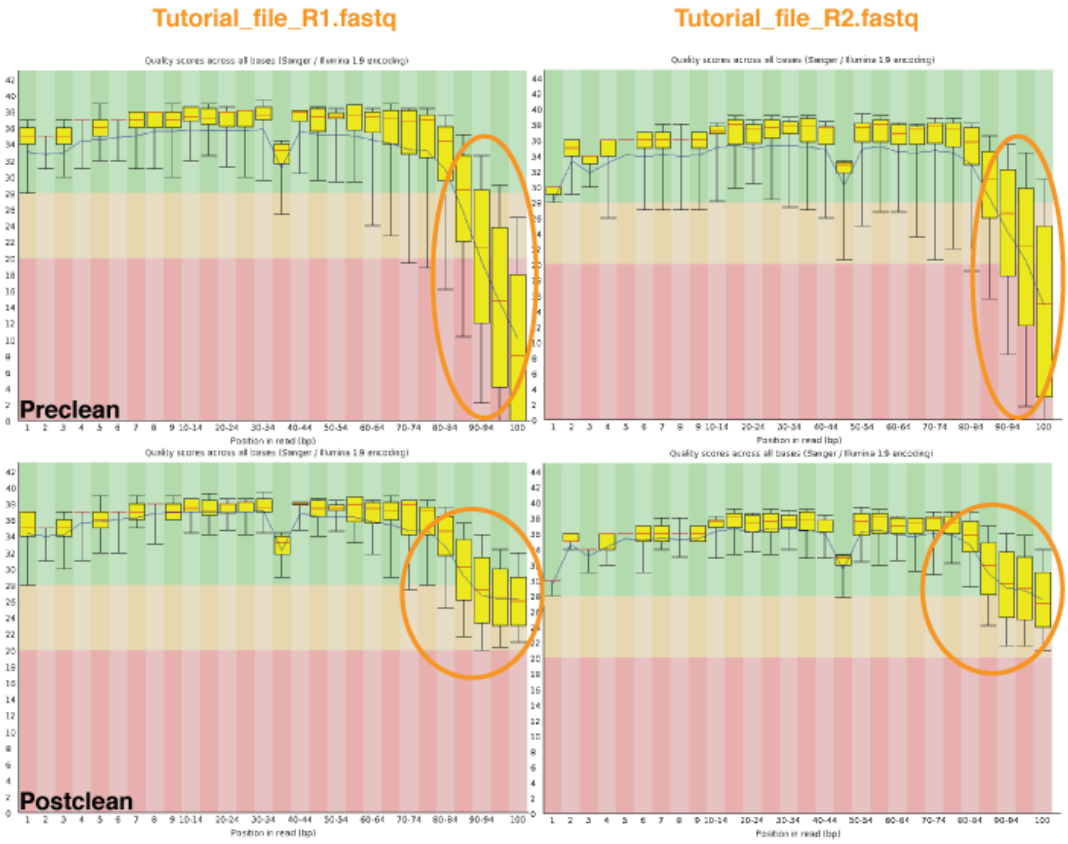


5 Review FastQC Results From Cleaned Datasets

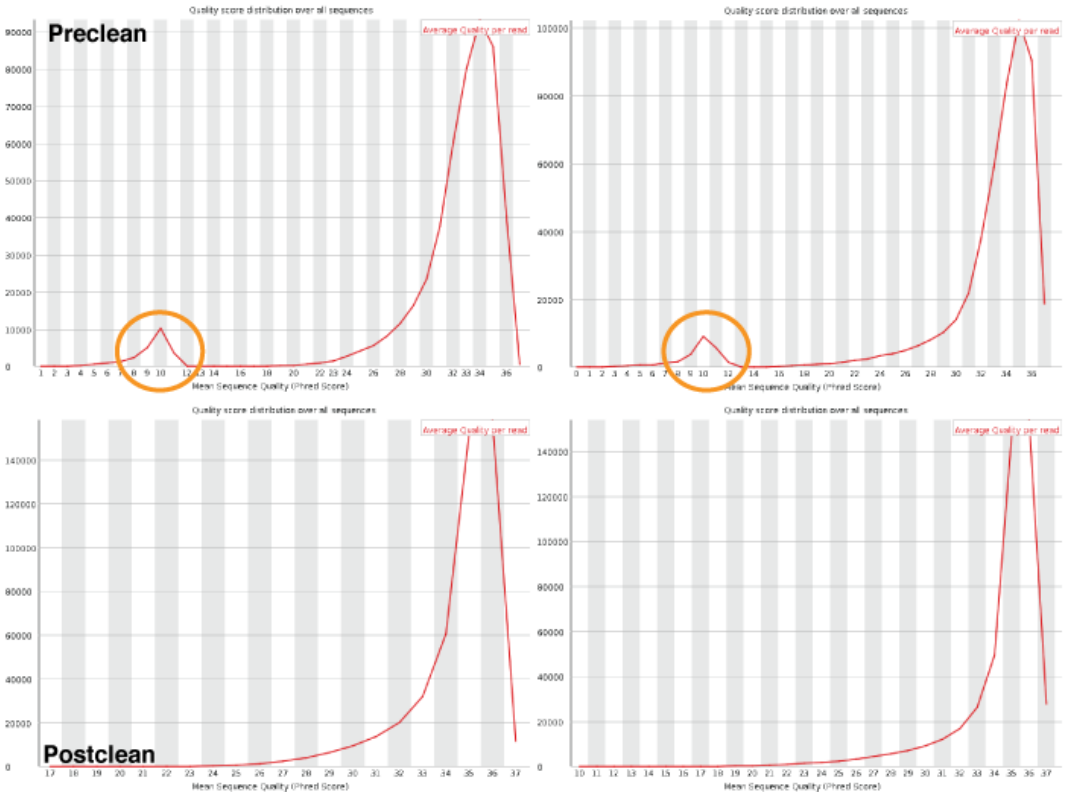
In this section we will compare the results from *FastQC* between the original FASTQ file and the quality and adapter trimmed FASTQ files. You should always examine the results post FASTQ file clean up before moving forward with more complex analysis. Here we will be showing both the results from the original FASTQ files as well as those from the quality and adapter trimmed FASTQ files.

a) Use *FastQC* to examine the quality statistics for the quality and adapter trimmed FASTQ files. Select the *Eye Icon* to view the results.

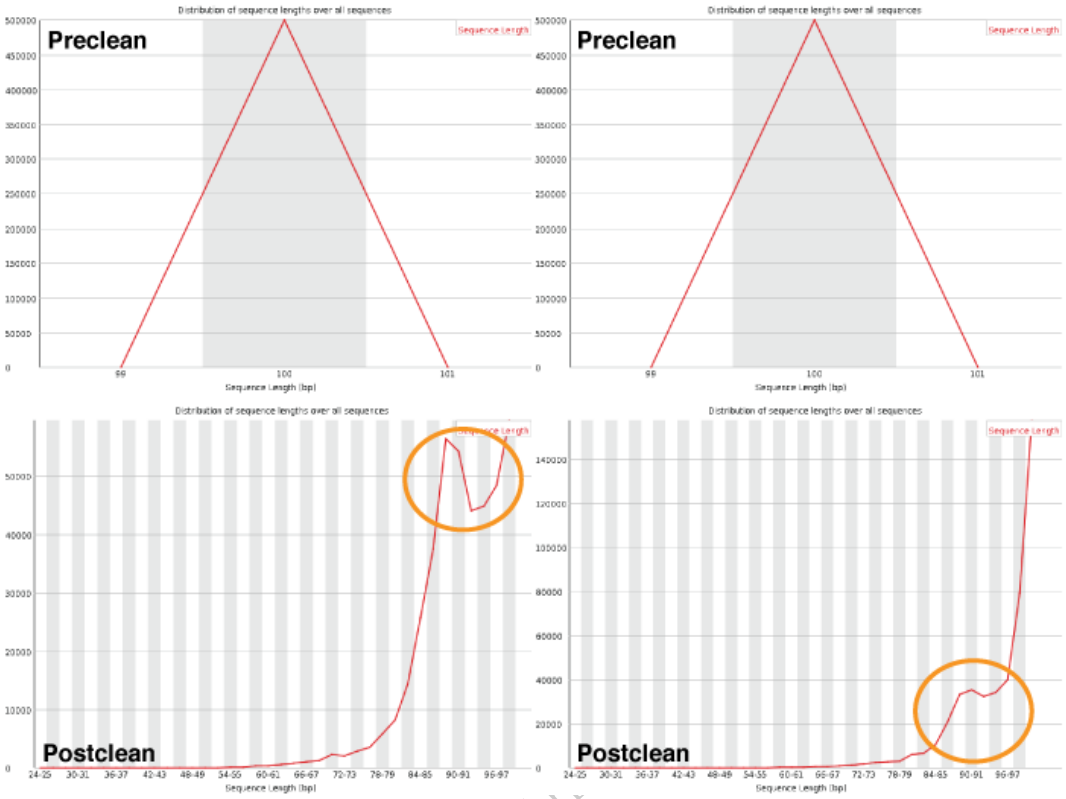
b) Scroll down to "Per base sequence quality". Note the improvement in the average quality of the read tails.



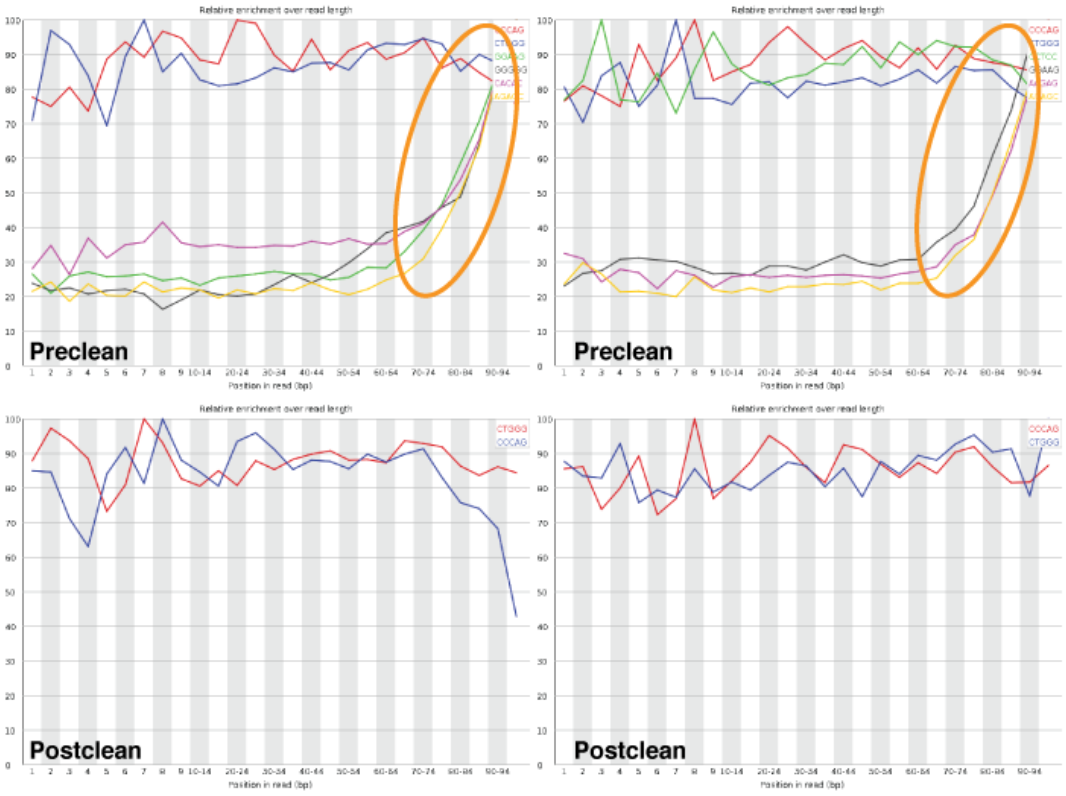
c) Scroll down to "Per sequence quality scores". Verify removal of low quality peak.



d) Scroll down to "Sequence Length Distribution". Note the degree of trimming that has occurred.



e) Scroll down to "Kmer Content". Note the absence of overrepresented tail k-mers.



6 Workflows

Creating a Workflow
The ability to create, reuse, share and publish workflows is on of Galaxy's largest strengths. Creating workflows allows you and anyone you want to collaborate with to exactly recreate analysis. You can think of workflows as your computational lab notebook, they are how you document your computational work. Workflows are also handy when you have to clean up your Galaxy space. Saving the raw input data and the workflow that leads to a final result allows you to delete the intermediate files yet retain the ability to recreate the entire analysis at any time. Workflows can be extracted from histories or created from scratch. Either method will result in a useable workflow so how you choose to build one is up to you.
Workflows are made up of connected tools, each tool is represented as a box and data moving from one tool to another is represented by the arrows. The inputs required for the tool can be found above the horizontal line in the box while the possible outputs are found below the line. Outputs from each tool can be saved and/or used as in the input for the next tool. Selecting the box will display the settings associated with the tool allowing you to preset parameters to reuse each time the workflow is run.

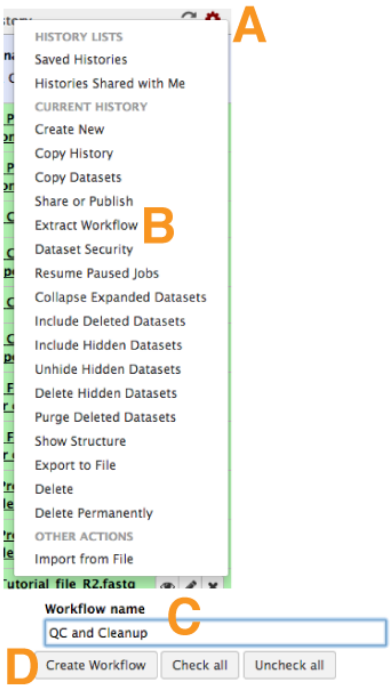
6.1 Extract Workflow from Current History

a) Select the *Gear Icon* from the top of the history pane.

b) Select *Extract Workflow* from the menu.

c) In *Workflow name* enter "QC and Cleanup".

d) Select *Create Workflow*.



6.2 View and Edit the Workflow

a) Select *Workflow* from the top bar.

b) Select the workflow that you just created and select *Edit* from the drop down menu.

c) The initial view of the workflow may be very messy. You can drag the boxes around on the screen to make the workflow easier to interpret. You can also move the blue box in the bottom right corner to view other sections of the workflow.

d) The workflow will follow the same logic as the history you created it from. Can you trace the steps you took for each initial FASTQ file though the QC and clean up process?

e) Select a *FastQC:Read QC* box which will open the *Details Pane* on the right. Is this for the Left or Right reads?

f) Select the *Input dataset* box that is attached (connected arrows) to the *FastQC:Read QC* box you just viewed.

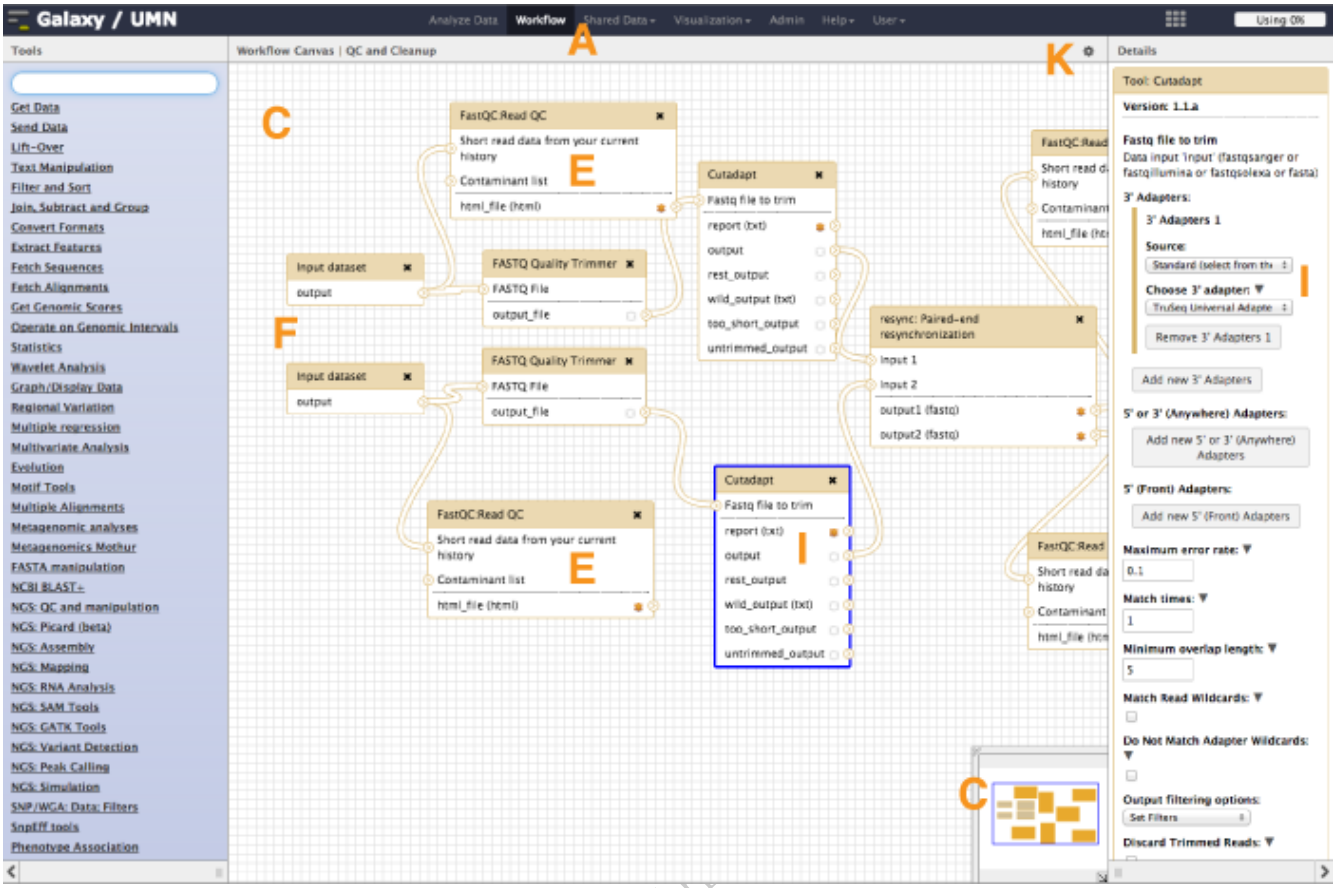
g) Label the *Input dataset* either Left or Right to correspond with the information from the *FastQC:Read QC* box you just viewed.

h) Do the same for the other *Input dataset*.

i) The next time you need to run QC and clean up FASTQ data you might need to use different adapter sequences. Selecting the *Cutadapt* box to view the *Details Pane*.

j) In the *Details Pane* for *Cutadapt* change the adapter sequence to "Set at runtime" using the *small arrow* next to *Choose 3' Adapter*. Do the same for the *Cutadapt* box associated with the other *Input dataset*.

k) Select the *Gear Icon* then *Save* from the menu.



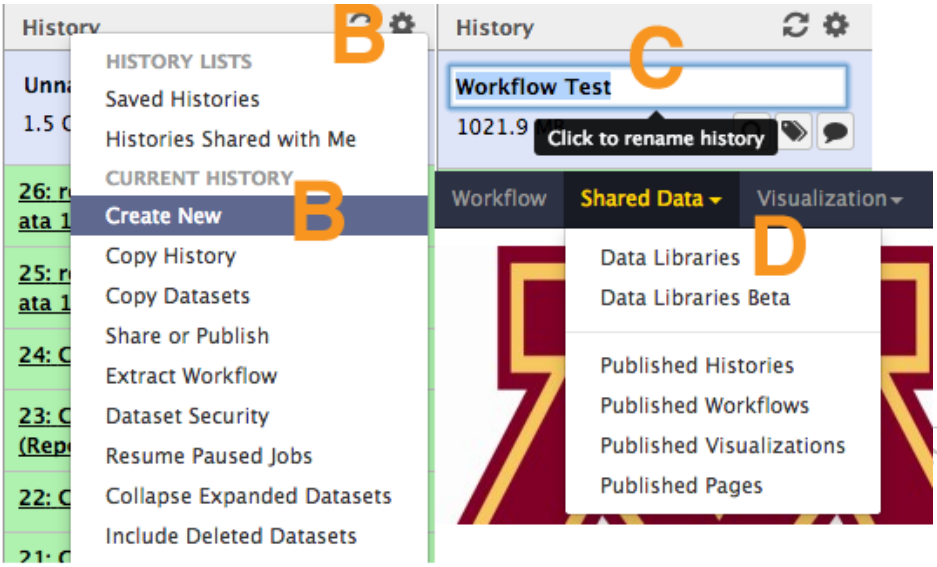
6.3 Running a Workflow

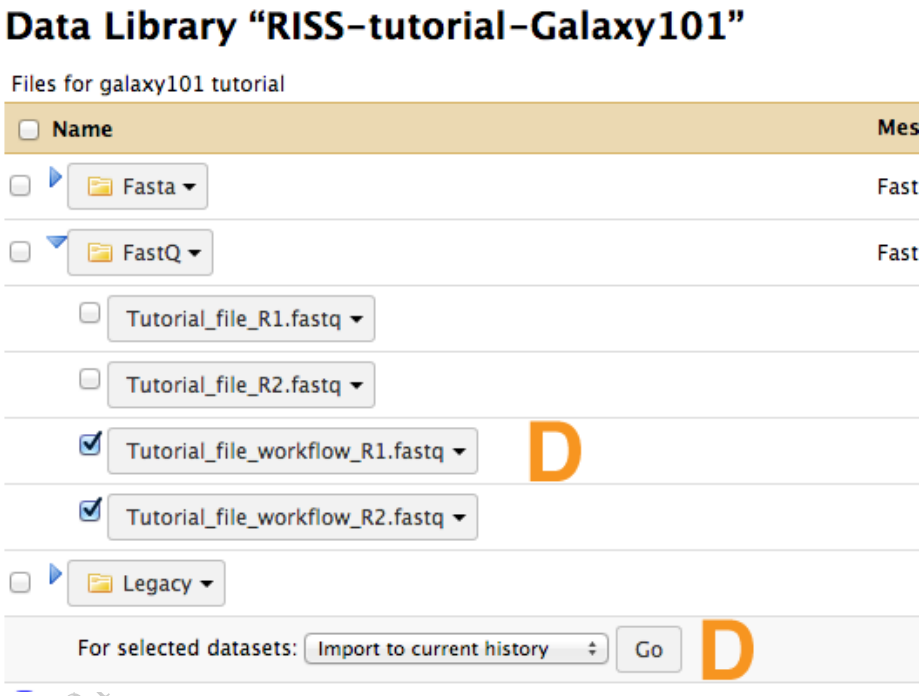
a) Select *Analyze Data* from the top bar to return to the main Galaxy screen.

b) Create a new history by selecting the *Gear Icon* then *Create New* from the menu.

c) Name the history "Workflow Test"

d) Import "Tutorial_file_workflow_R1.fastq" and "Tutorial_file_workflow_R2.fastq" into the current "Workflow Test" history from the data library(Section 2.2. Don't forget to set the file attributes (Section 2.3).





e) Select *Workflow* from the top bar to display your saved workflow from the data library.

f) Select the *QC and Cleanup* workflow you just created then select *Run* from the drop down menu.

g) In the Center Pane select the "_R1.fastq" file for the *Left Read Input* using the drop down menu.

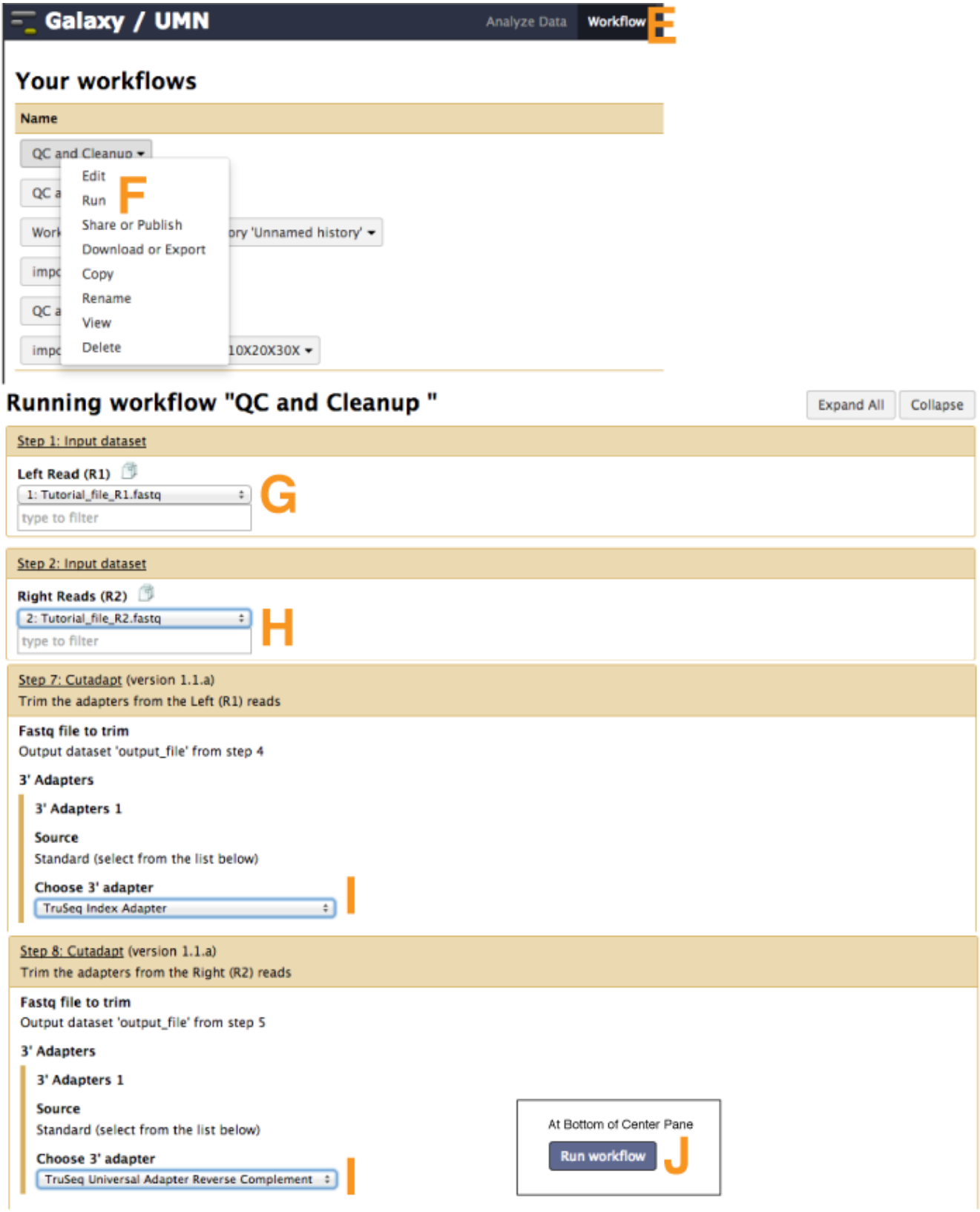
h) Select the "_R2.fastq" file for the *Right Read Input* using the drop down menu.

i) Scroll down the the *Cutadapt* steps and select the appropriate adapters.

j) Scroll to the bottom of the main view and select *Run workflow*

k) Select *Analyze Data* in the top bar to return to the main Galaxy view.

l) You will be able to watch the progress of the workflow in the History Pane.



7 Sharing Workflows and Histories

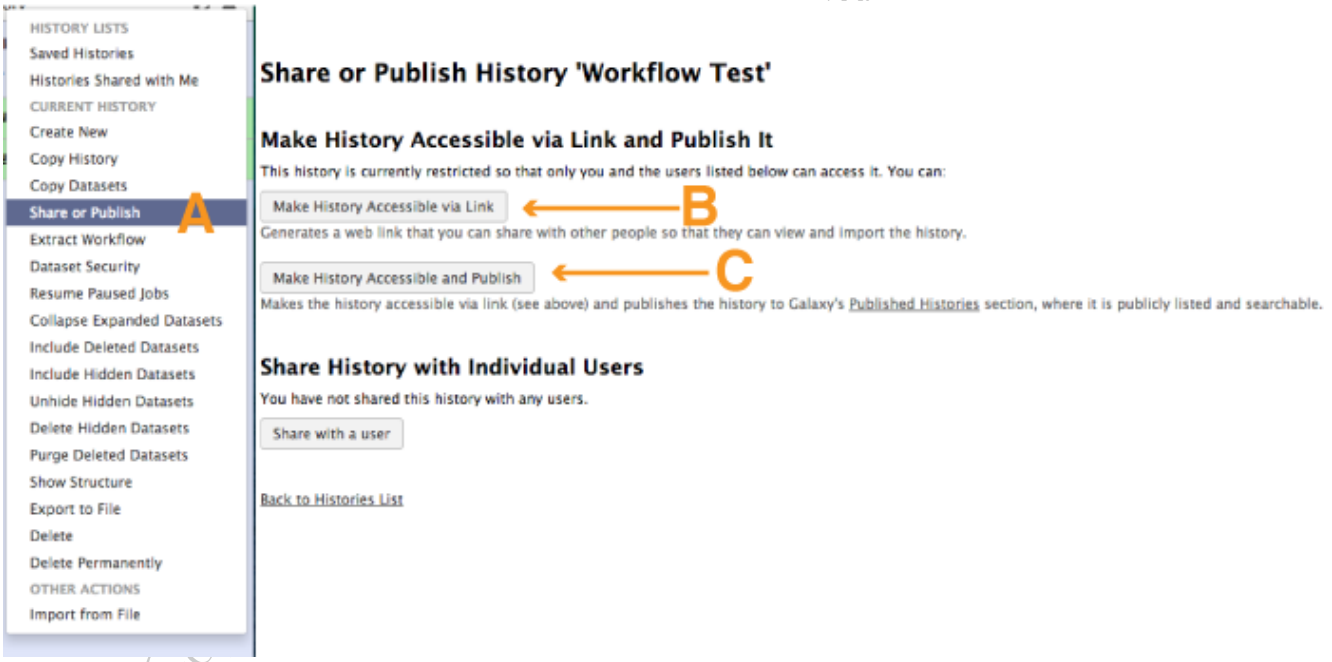
It is possible to share workflow and histories with other Galaxy users. This allows you to share data, results and methods with collaborators or anyone who might want to recreate your methodology. Galaxy histories and workflows can be shared via a link or they can be saved as stand alone files that can then be uploaded to any Galaxy instance.

7.1 Share a History

a) To share your current history select the *Gear Icon* then *Share or Publish*

b) To share the history though a web link select *Make History Accessible via Link*. You can share this link with anyone who has access to Galaxy at MSI allowing them to view the history and the data in it.

c) *Make History Accessible and Publish* will also create a link to the history but it will also publish the history making it public to anyone with access to Galaxy at MSI under the Shared Data tab.



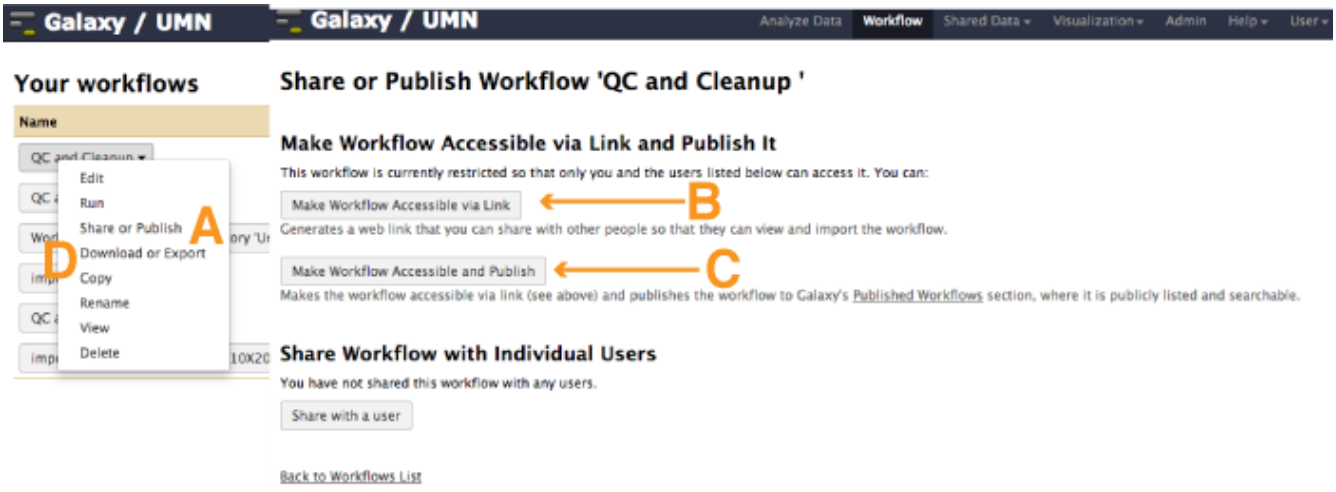
7.2 Share a Workflow

a) When you select a workflow from the list one of the options is *Share or Publish*

b) To share the workflow though a web link select *Make Workflow Accessible via Link*. You can share this link with anyone who has access to Galaxy at MSI allowing them to view and use the workflow.

c) *Make Workflow Accessible and Publish* also creates a link to the workflow but it will also publish the workflow making it public to anyone with access to Galaxy at MSI under the Shared Data tab.

d) You can also download a workflow to be imported into another Galaxy instance or to be archived by selecting *Download or Export*.



8 Cleaning Up Histories: Deleting Data From Galaxy

Galaxy is a shared resource so the amount of data you and your group can have in Galaxy is limited. We estimate that even a relatively simple RNAN-seq analysis will use 4-5 times the storage of the raw sequencing files. Many of these files are intermediate and can be discarded once the analysis is complete. Also, since Galaxy allows you to create workflows it is easy to recreate intermediate files if they are needed later. It is good practice to extract workflows from histories then discard the histories once you have completed the analyses.
Your groups current Galaxy usage is displayed in the top bar on the far right. You can view the size of your different Galaxy histories when you view you saved Galaxy histories using the *Gear Icon* and below the name of your current history. If you delete a Galaxy history before you extract a workflow you will not be able to do so later.

8.1 Deleting Intermediate Files and Histories from Galaxy

a) Select *Analyze Data* in the top bar to get to the main Galaxy view.

b) Select the *Gear Icon* and then *Saved Histories* from the menu.

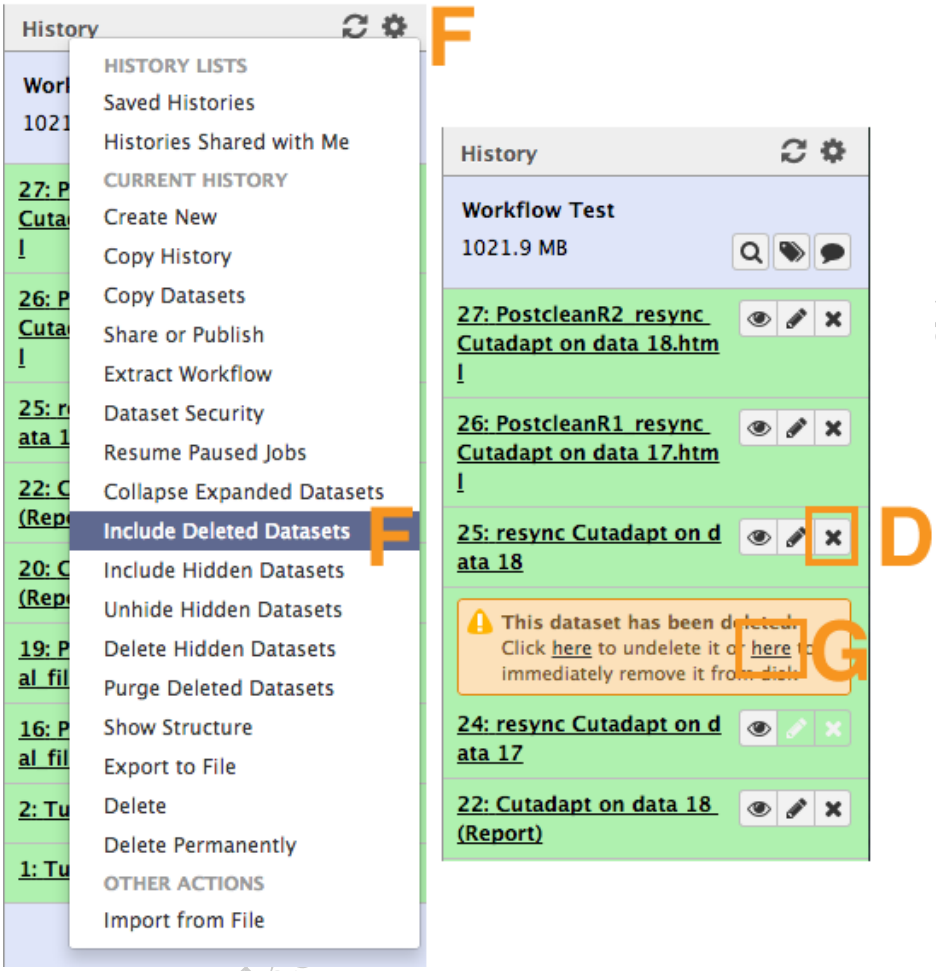
c) Select the history you created when you tested your workflow then select *Switch* to open the history in the Galaxy History Pane.

d) To delete specific pieces of data from a Galaxy history you can select the *X*.

e) Notice that when the data set is deleted that the size of the history does not change. This is because Galaxy has a recycling bin type function.

f) To permanently delete a dataset first unhide the hidden datasets by selecting the *Gear Icon* then *Include Deleted Datasets*.

g) Select the *here* link displayed in the history pane for the data you would like to delete. This will actually reduce the size of the history.



h) You can delete an entire history from the same page where you can view your saved histories.

i) Select the *Gear Icon* then *Saved Histories*

j) Select the history that you want to delete. *Delete Permanently* will remove the history immediately while *Delete* will place the history in the recycling bin.

k) While data can be restored from the recycle bin MSI will clear out the recycle bin monthly so if you choose to delete a history you should just *Delete Permanently*.

Saved Histories

search history names and tags

Advanced Search

	Name	Datasets	Tags	Sharing	Size on Disk	Created	Last Updated ↑	Status
☐	Workflow Test	10	0 Tags		1021.9 MB	~1 day ago	~1 day ago	current history
☐	Unnamed history	12	0 Tags		1.5 GB	Mar 26, 2014	~2 days ago	
☐	QC and Clean Fresh	10	0 Tags		4.7 GB	Mar 26, 2014	Mar 26, 2014	

For 0 selected histories: Rename | Delete | Delete Permanently | Undelete

Histories that have been deleted for more than a time period specified by the Galaxy administrator(s) may be permanently deleted.