Galaxy RNA-Seq Analysis: *H. sapiens*

Tutorial
Research Informatics Support Systems
Minnesota Supercomputing Institute
University of Minnesota
Version 2
10/8/2013

1 Introduction

1.1 Scope of this tutorial

This is a practical, hands-on tutorial designed to give participants experience with RNA-Seq data analysis using Tophat, Cufflinks, and CummRbund in Galaxy. The analysis in this tutorial is typical of experiments in eukaryotic species with high-quality genomes and genome annotation available. Participants are expected to be familiar with next-generation sequence data, basic theory of RNA-Seq, and Galaxy. Participants do not need previous experience with Tophat, Cufflinks, or CummeRbund.

1.2 Reference materials

RNA-Seq Lecture PDFs on MSI website: www.msi.umn.edu/content/bioinformatics-analysis
Galaxy 101: NGS data analysis hands-on tutorial:
www.msi.umn.edu/content/bioinformatics-analysis
Tophat manual: tophat.cbcb.umd.edu/manual.html
Cufflinks manual: cufflinks.cbcb.umd.edu/manual.html
CummeRbund manual: compbio.mit.edu/cummeRbund

1.3 Outline of tutorial

2 Starting Galaxy

★Tutorial Dataset (Sect 2.2 page 6)
This tutorial will identify genes whose expression levels differ between skeletal muscle tissue and heart muscle tissue. The sample dataset used in this tutorial was created from the heart and skeletal muscle samples from the Illumina Bodymap 2.0 Project (**http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30611**). The single heart and skeletal muscle samples were split into three subsamples, and the reads mapping to a 5MB region near the distal end of chromosome 19 were extracted along with some unmapped reads. Each fastq file contains about 50,000 50 base-pair paired-end reads.

NOTE: This dataset was chosen to allow for fast processing and response times in a classroom setting where dozens of people will be submitting jobs at once to the server. It is not ideal due to the small sample sizes (leading to atypical-looking graphs in some cases and poor statistics) and lack of real biological replicates (resulting in unrealistically-good sample separation).

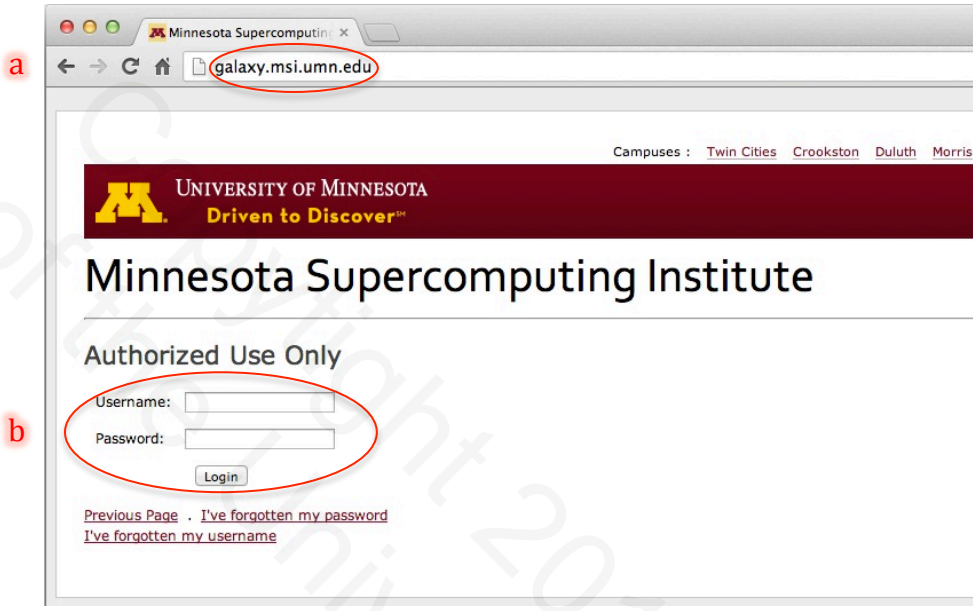
★GTF Files (Sect 2.3 page 7)
A GTF file identifies the genomic locations of genes and their exons. If a GTF file for your organism is not listed send a request to MSI, or find one online at sites such as www.ensembl.org/info/data/ftp/index.html, genome.ucsc.edu/cgi-bin/hgTables?command=start, or NCBI. The GTF files provided in the Illumina iGenomes collection (cufflinks.cbcb.umd.edu/igenomes.html) have been specially modified for maximum compatibility with the Cufflinks and Cuffdiff programs.

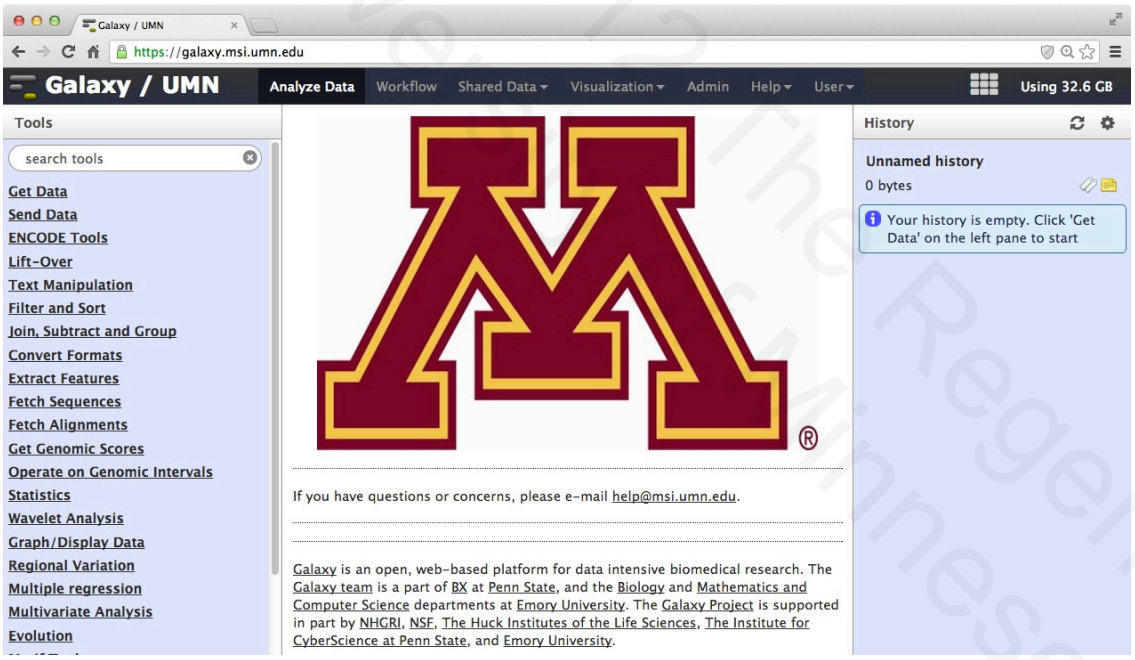
★Quality Control (Sect 2.5 page 9)
It is important to always verify the integrity of a dataset before starting to analyze it. Quantifying dataset quality may uncover problems that might otherwise go undetected. Data quality problems such as sequencing adaptor contamination or low read quality require trimming and filtering not covered in this tutorial. See the Galaxy 101 tutorial handout on the MSI website for detailed instructions on how to clean up a low quality dataset:
www.msi.umn.edu/content/bioinformatics-analysis

The graphs generated in this tutorial are not entirely typical due to the small sample datasets used. See the typical output from a good Illumina dataset:
www.bioinformatics.babraham.ac.uk/projects/fastqc/good_sequence_short_fastqc/fastqc_report.html, and a poor Illumina dataset:
www.bioinformatics.babraham.ac.uk/projects/fastqc/bad_sequence_fastqc/fastqc_report.html.
For more information about interpreting FastQC output refer to the RISS RNA-Seq Lecture 1 tutorial handout: www.msi.umn.edu/content/bioinformatics-analysis

2.1 Accessing Galaxy

a) Open a web browser and navigate to MSI Galaxy website galaxy.msi.umn.edu
b) Log in with your MSI username and password





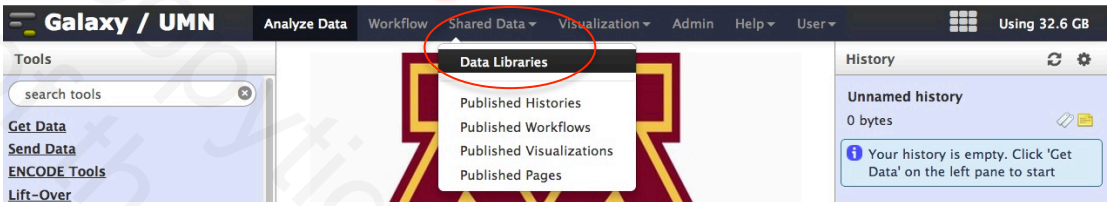
Tools pane Center pane History pane

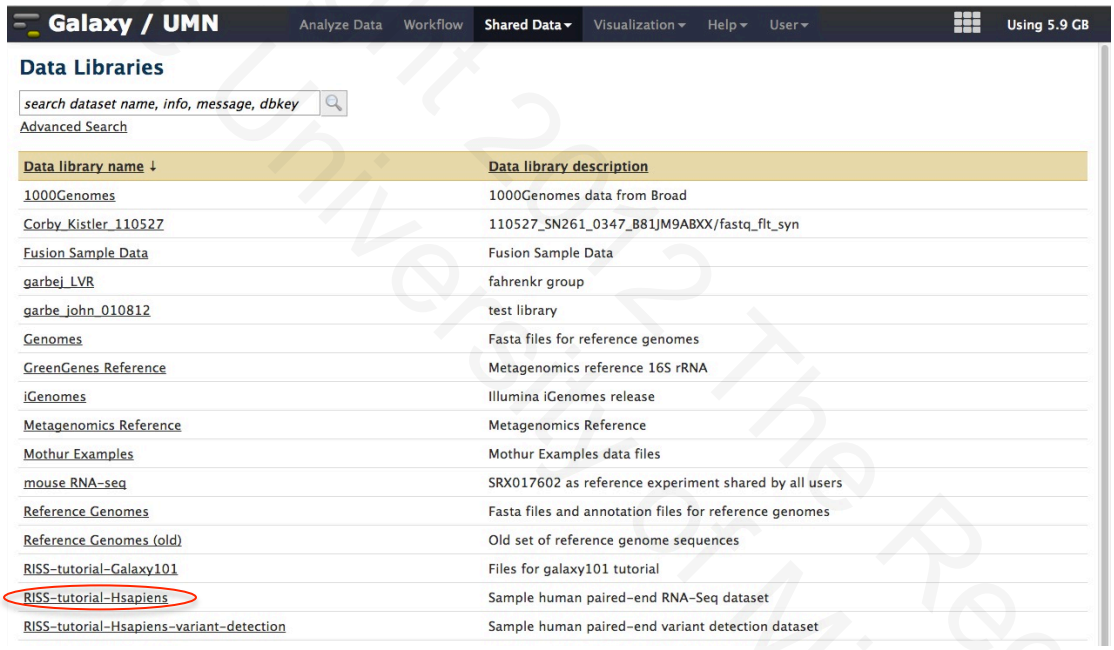
2.2 Import Fastq files for one sample into current history

★Tutorial Dataset

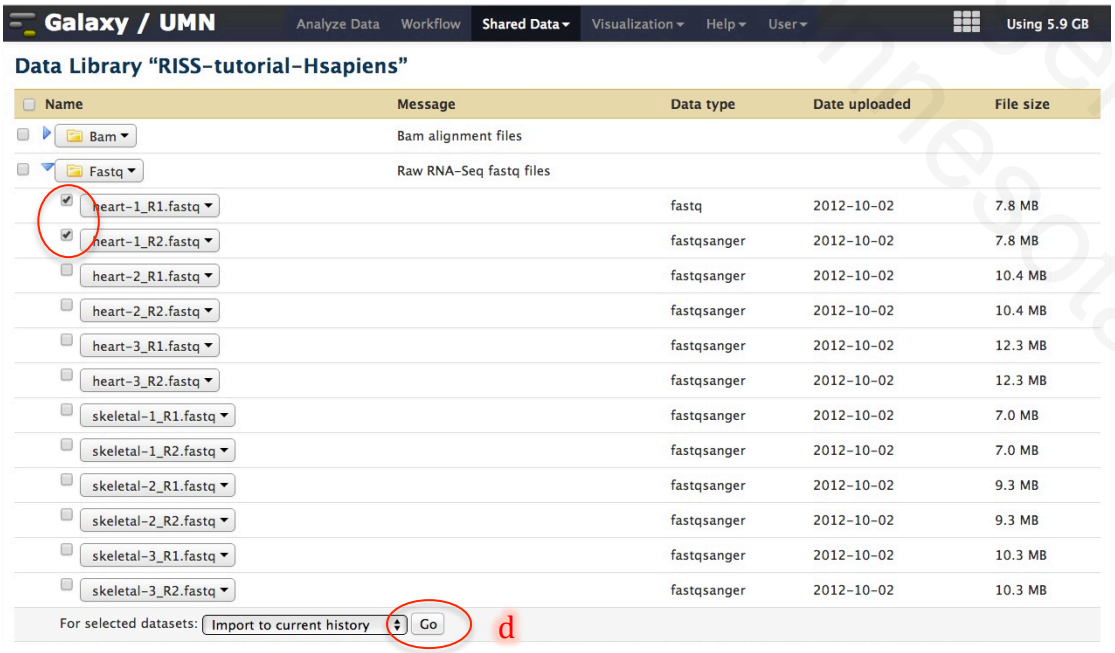
a) At the top of the screen select “Shared Data -> Data Libraries”
b) Select “RISS-tutorial-Hsapiens” from the list of data libraries
c) Expand the “Fastq” folder and check the boxes next to the first two files
d) Near the bottom of the page click the “Go” button to import the selected datasets to the current history

a





b



c

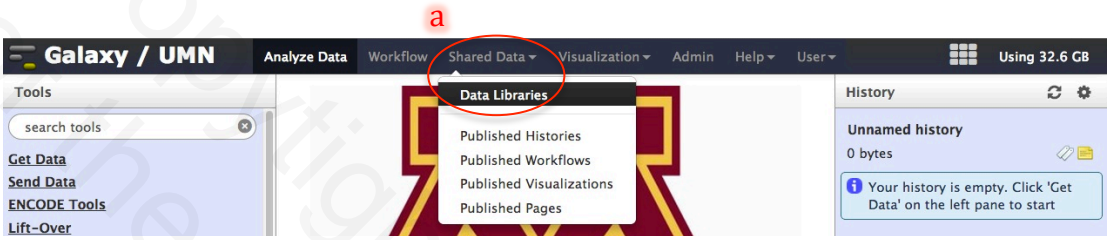
d

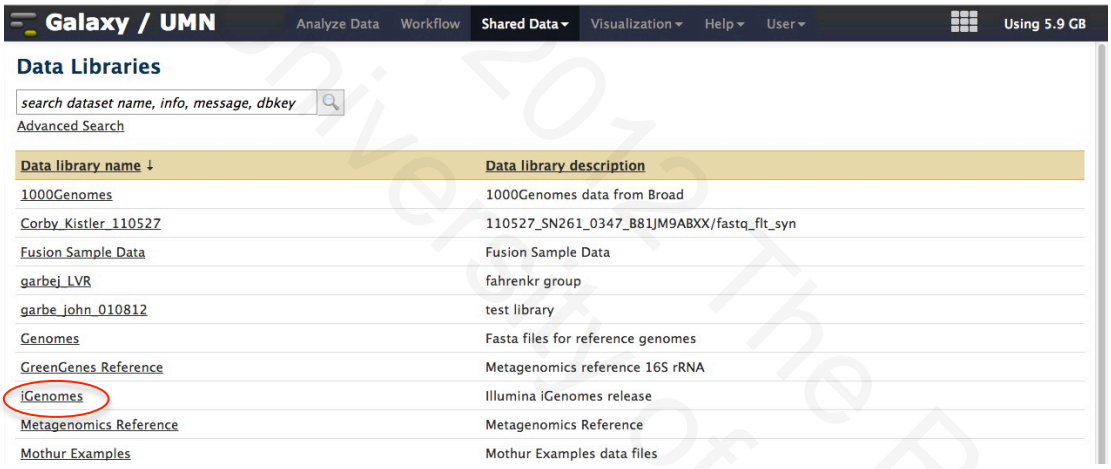
2.3 Import the GTF file from the iGenomes data library

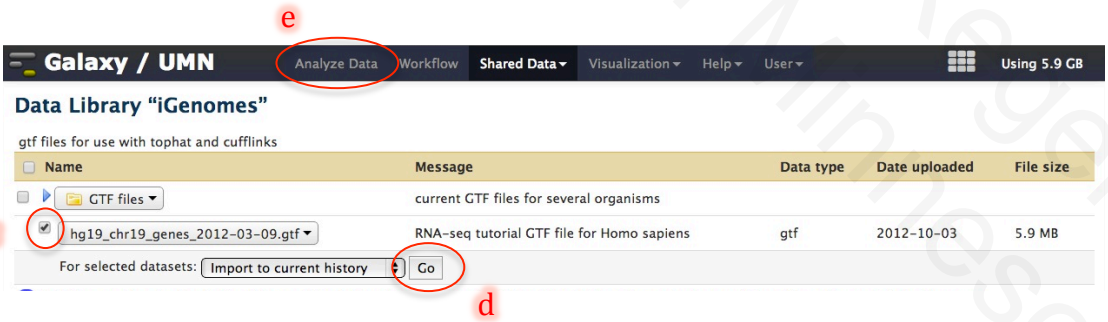
★GTF Files

a) At the top of the screen select “Shared Data -> Data Libraries”
b) Select “iGenomes” from the list of data libraries
c) Check the box next to the “hg19_chr19_genes_2012-03-09.gtf” file
d) Near the bottom of the page click the “Go” button to import the selected datasets to the current history
e) At the top of the screen select “Analyze Data” to return to your current history

---



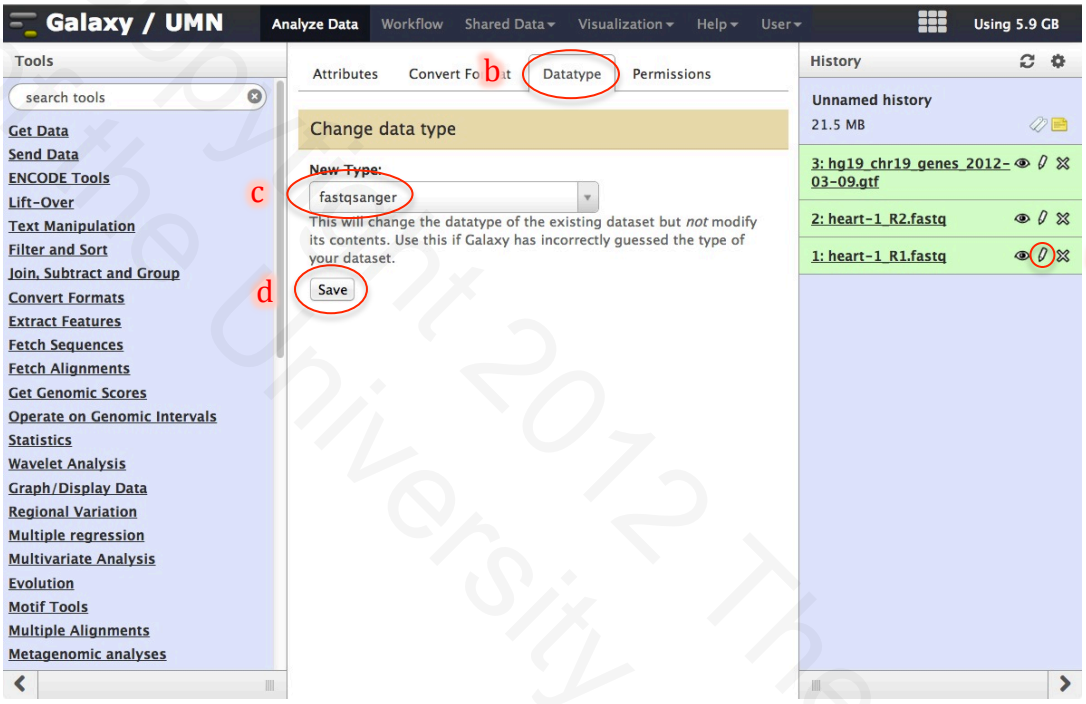




2.4 Set file attributes

a) In the history pane click on the pencil icon next to the heart-1_R1.fastq file
b) Click the Datatype tab
c) Enter “fastqsanger” in the “New Type” box. A list of available data types will appear as you type.
d) Click save

> For a real dataset you would need to repeat this step on the R2 fastq file



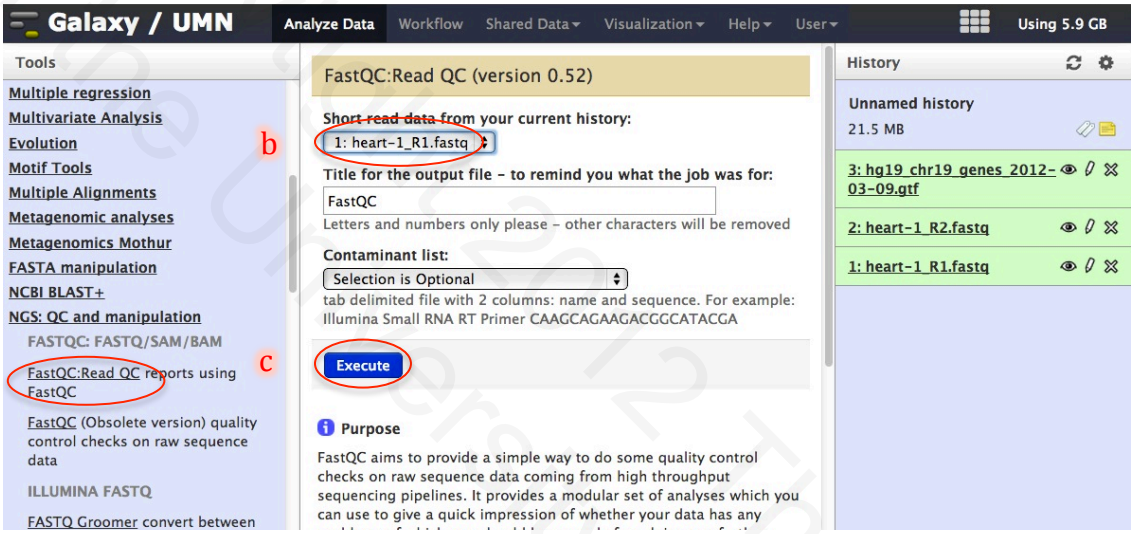
2.5 Run FastQC

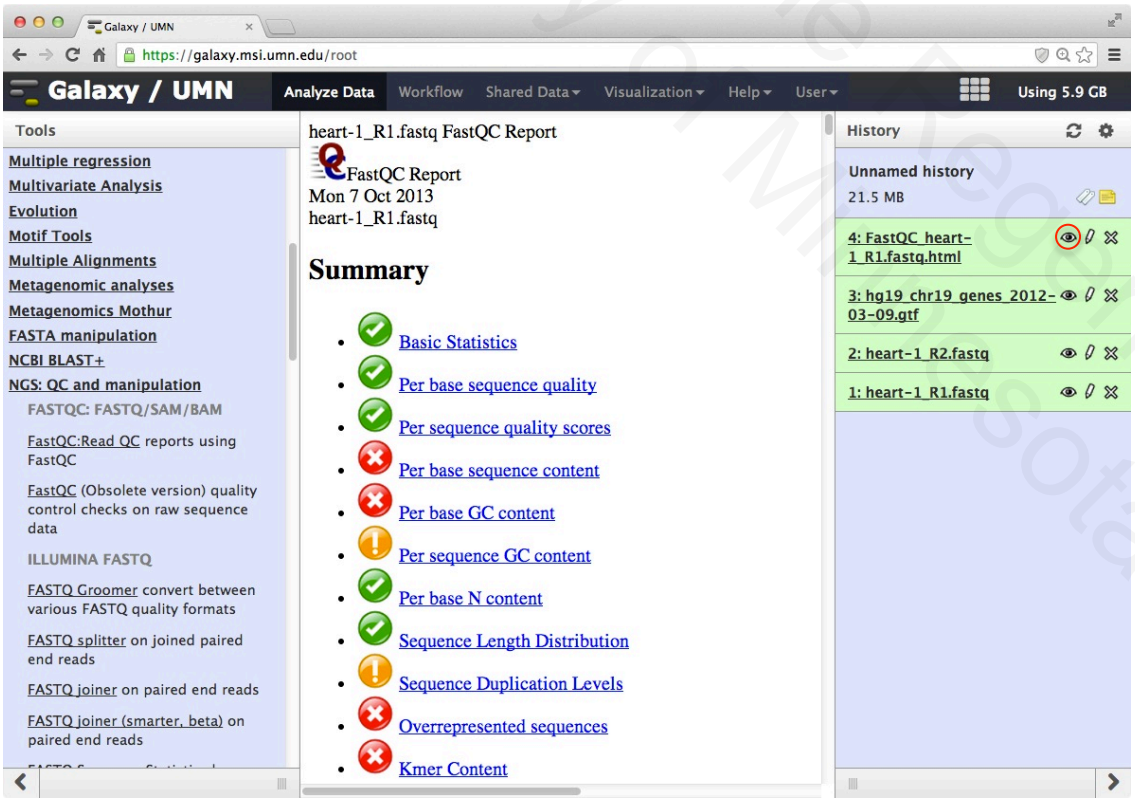
★Quality Control

a) Load the FastQC tool from the tool pane: "NGS: QC and manipulation -> FastQC: Read QC"
b) Set the input file: select "heart-1_R1.fastq" from the dropdown menu under "Short read data from your current history"
c) Click "Execute"
d) When fastqc has finished running, click on the eye on the FastQC output file to display the file in the center pane

⚠ For a real dataset you would need to repeat this step on the R2 fastq file

⚠ See the Galaxy 101 tutorial handout for detailed instructions on how to clean up a low quality dataset: www.msi.umn.edu/content/bioinformatics-analysis

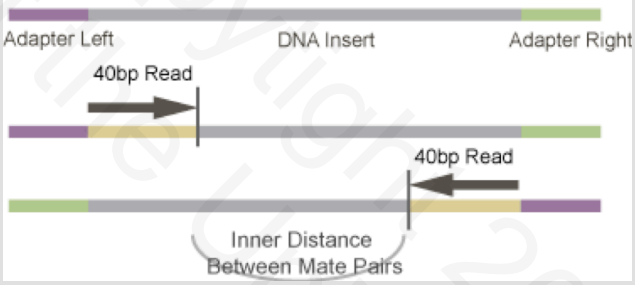




3 Mapping with Tophat

★Reference Genomes (Sect 3.1 page 11)
It is important that the reference genome you align against is generated from the same reference genome as the GTF you are using because the chromosome names and coordinates used in the GTF file must be the same as those used in the database. See www.msi.umn.edu/content/reference-genomes for full details about the reference genomes available in Galaxy. If the reference genome for your organism is not listed email a request to MSI to have it added.

★Mean Inner Distance – Part I (Sect 3.1 page 11)
This is the expected (mean) inner distance between mate pairs. For example, the UMGC's default fragment selection size is 200, so 200 – (2 * read length) is a good value to use for this parameter. We will determine the exact fragment length in the next section.

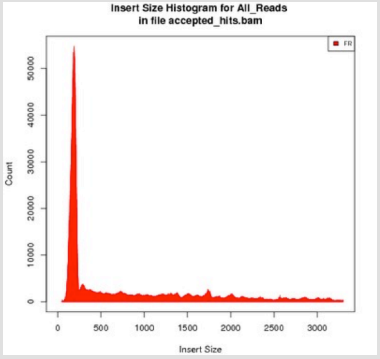


★Junctions (Sect 3.1 page 11)
Tophat can attempt to identify exon-exon splice junctions solely using your dataset, or you may supply a set of gene model annotations as a GTF or GFF file. In this tutorial we will provide a GTF annotation file because the human genome is well annotated.

★Advanced Tophat Parameters (Sect 3.1 page 11)
See the RNA-Seq Lecture 2 handout for more detail on setting parameters properly for other organisms: www.msi.umn.edu/content/bioinformatics-analysis

★Mean Inner Distance – Part II (Sect 3.2 page 12)
It is important that the mean inner distance Tophat parameter is set correctly in order to get the best mapping results. The actual average fragment size for each sample can be determined by running Tophat with an estimated inner distance and then calculating the true value from the mapped reads. Rerunning Tophat with the true value will give improved results.

★Insert Size Histogram (Sect 3.2 page 12)
The insert size histogram generated from this sample dataset is noisier than a typical histogram, shown here:



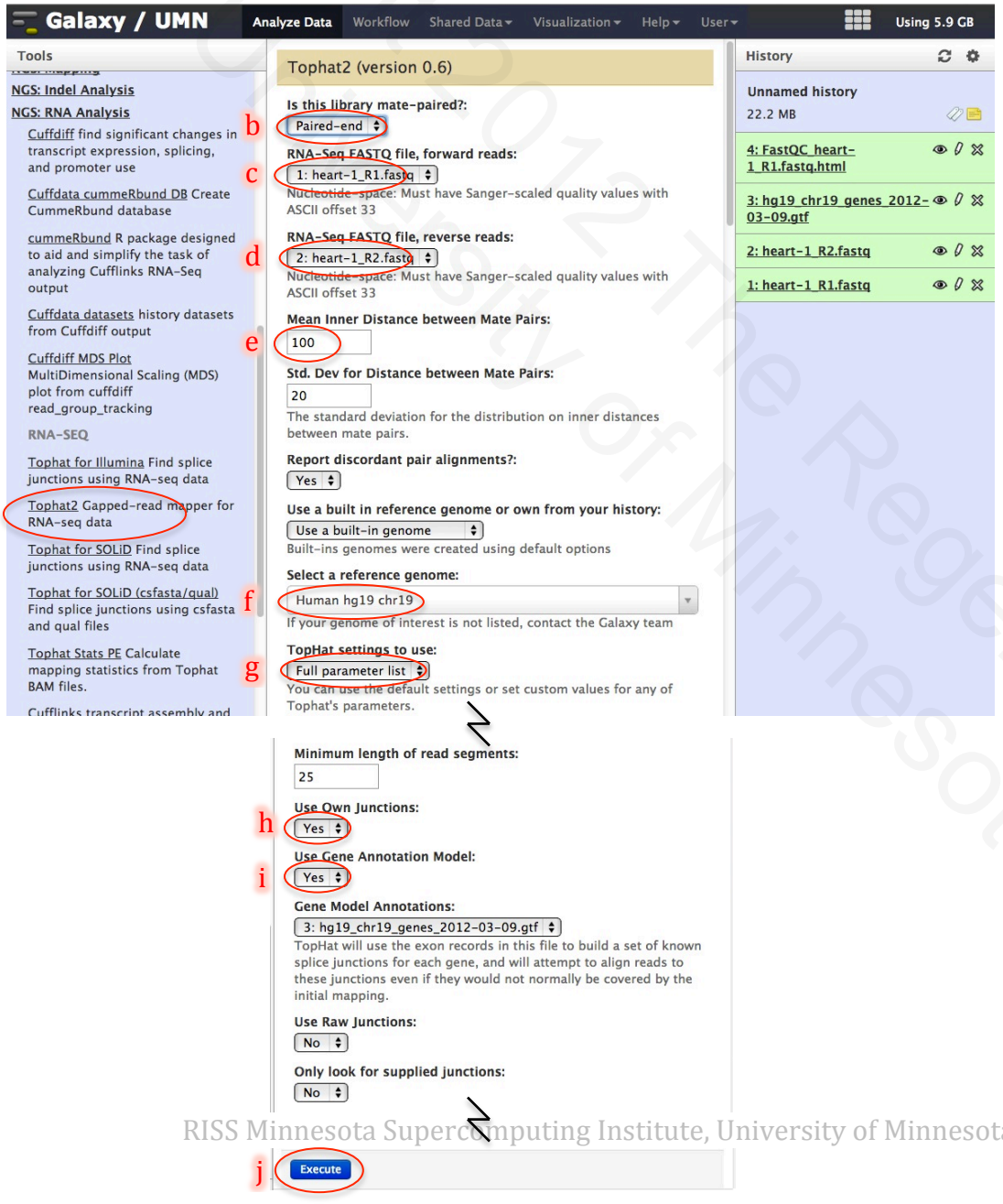
★Mapping Statistics (Sect 3.4 page 14)
It is important to determine how well the RNA-Seq reads align to the reference genome. Low mapping rates require further investigation to determine the cause.

3.1 Initial Tophat run

★Reference Genomes
★Mean Inner Distance – Part I
★Junctions
★Advanced Tophat Parameters

a) Load the Tophat2 tool from the tool pane: "NGS: RNA Analysis -> Tophat2"
b) Is this library mate-paired -> Paired-end
c) RNA-Seq FASTQ file, forward reads -> heart-1_R1.fastq
d) RNA-Seq FASTQ file, reverse reads -> heart-1_R2.fastq
e) Mean Inner Distance between Mate Pairs -> 100
f) Select a reference genome -> Human hg19 chr19
g) TopHat settings to use -> Full parameter list
h) Use Own Junctions -> Yes
i) Use Gene Annotation Model -> Yes
j) Click "Execute" to submit the job

> ⚠ Only files of type "fastqsanger" will appear in the dropdown list. If your fastq file isn't shown the file type is set incorrectly. See step 2.4



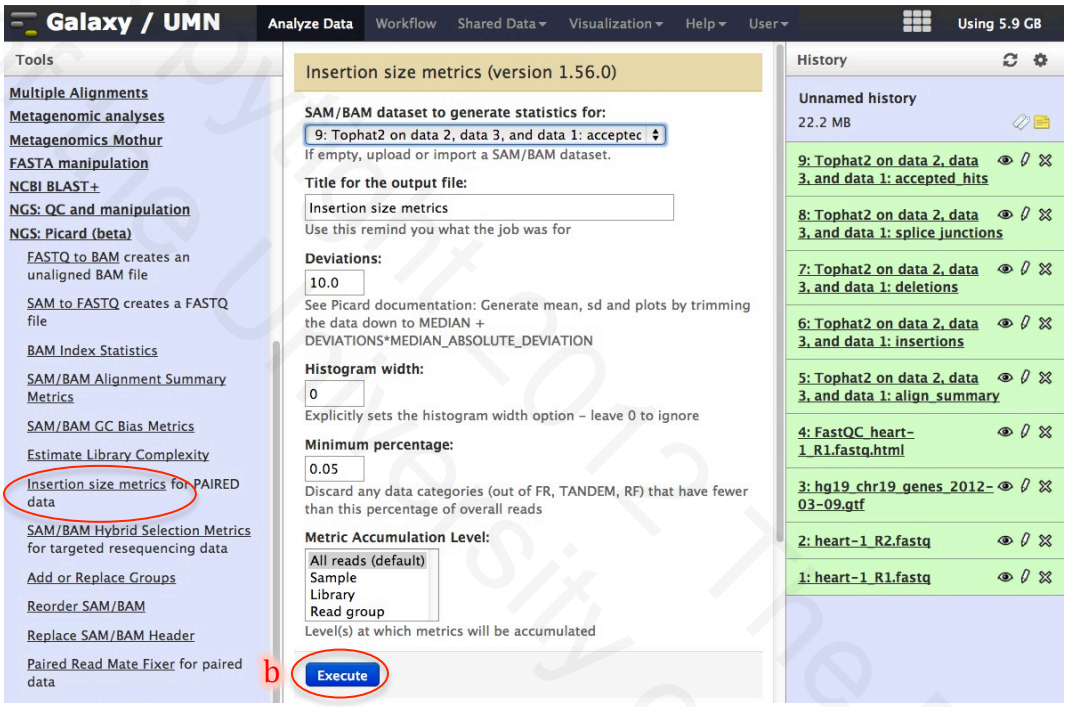
3.2 Determine insert size

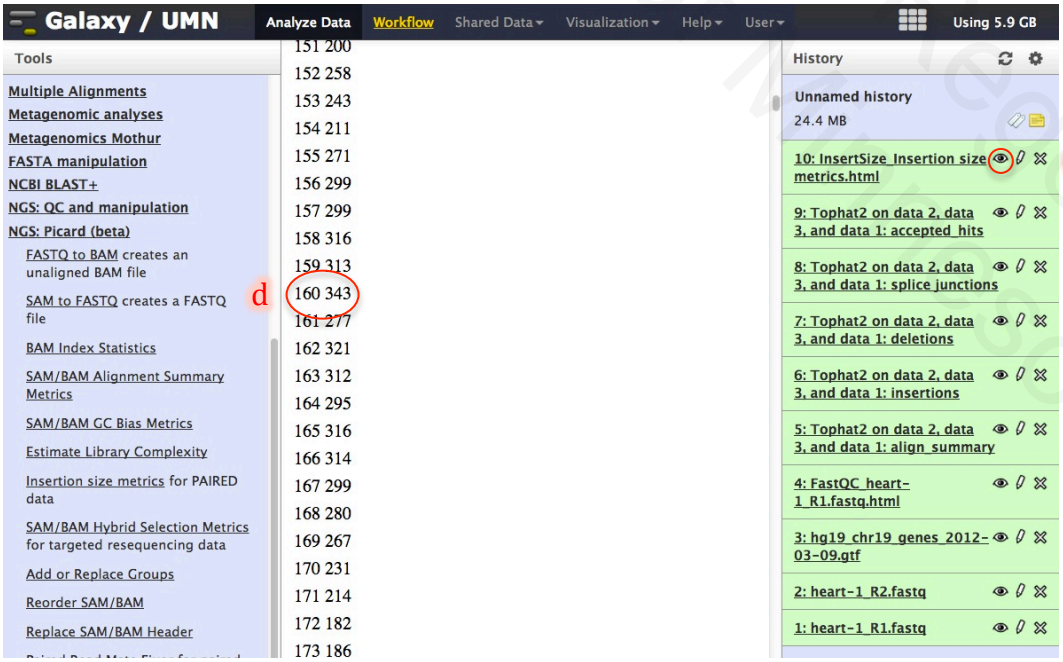
★Mean Inner Distance – Part II

★Insert Size Histogram

a) Load the insert size tool "NGS: Picard (beta) -> Insertion size metrics"
b) Click Execute
c) Click on the "eye" icon next to the output file in the history pane to view the output in the central pane
d) Identify the mode (highest frequency) insert size from the program output

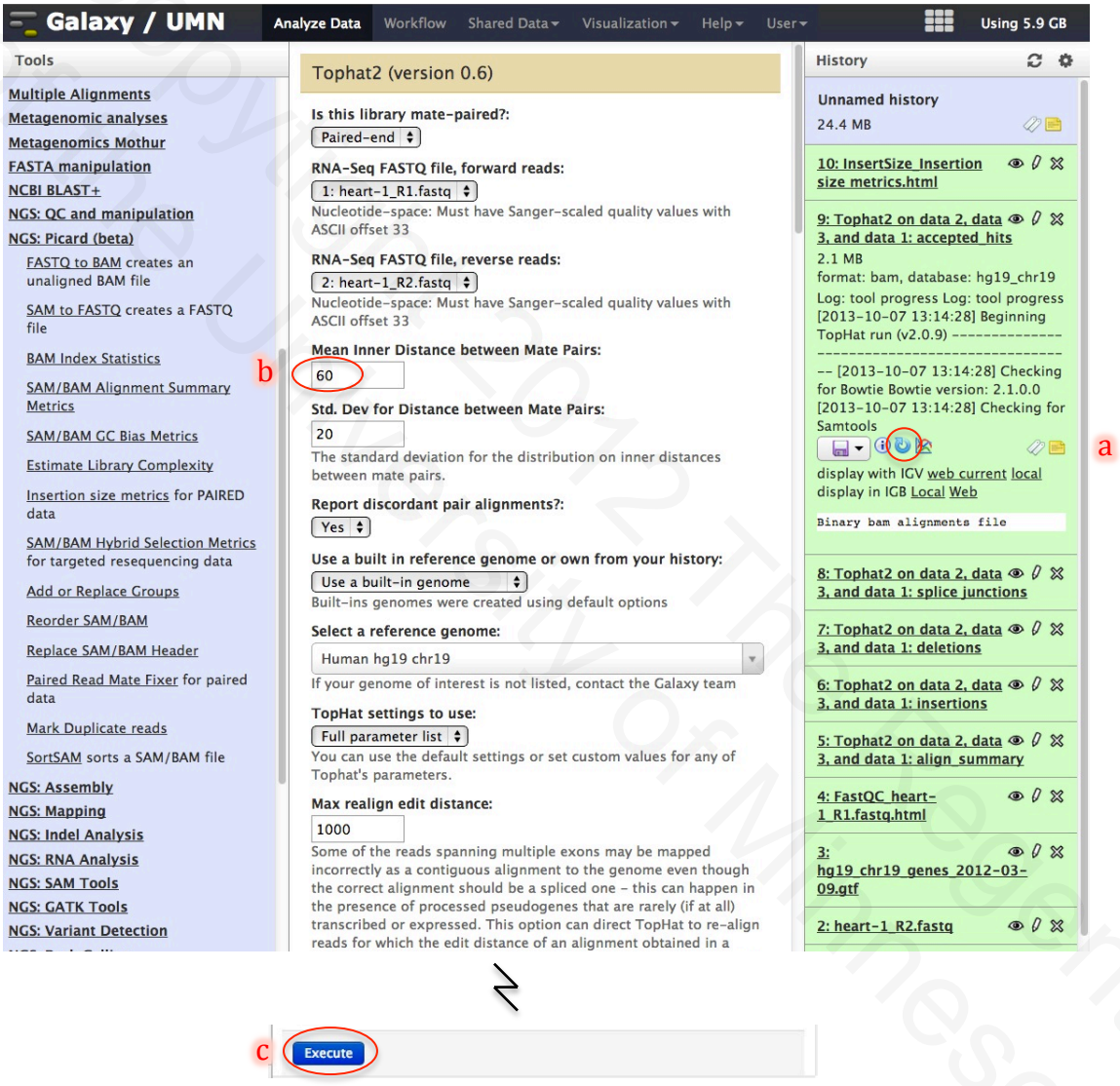
---





3.3 Rerun Tophat with correct insert size

a) Click on the name of any one of the Tophat2 output files in the history pane to expand it, and click on the circular blue arrow icon to display the Tophat2 tool in the central pane with the parameters preset from the last Tophat2 run
b) Change the "Mean Inner Distance between Mate Pairs" to the correct value: Picard value – (2 * read length) = 160 – (2 * 50) = 60
c) Click "Execute" to submit the job

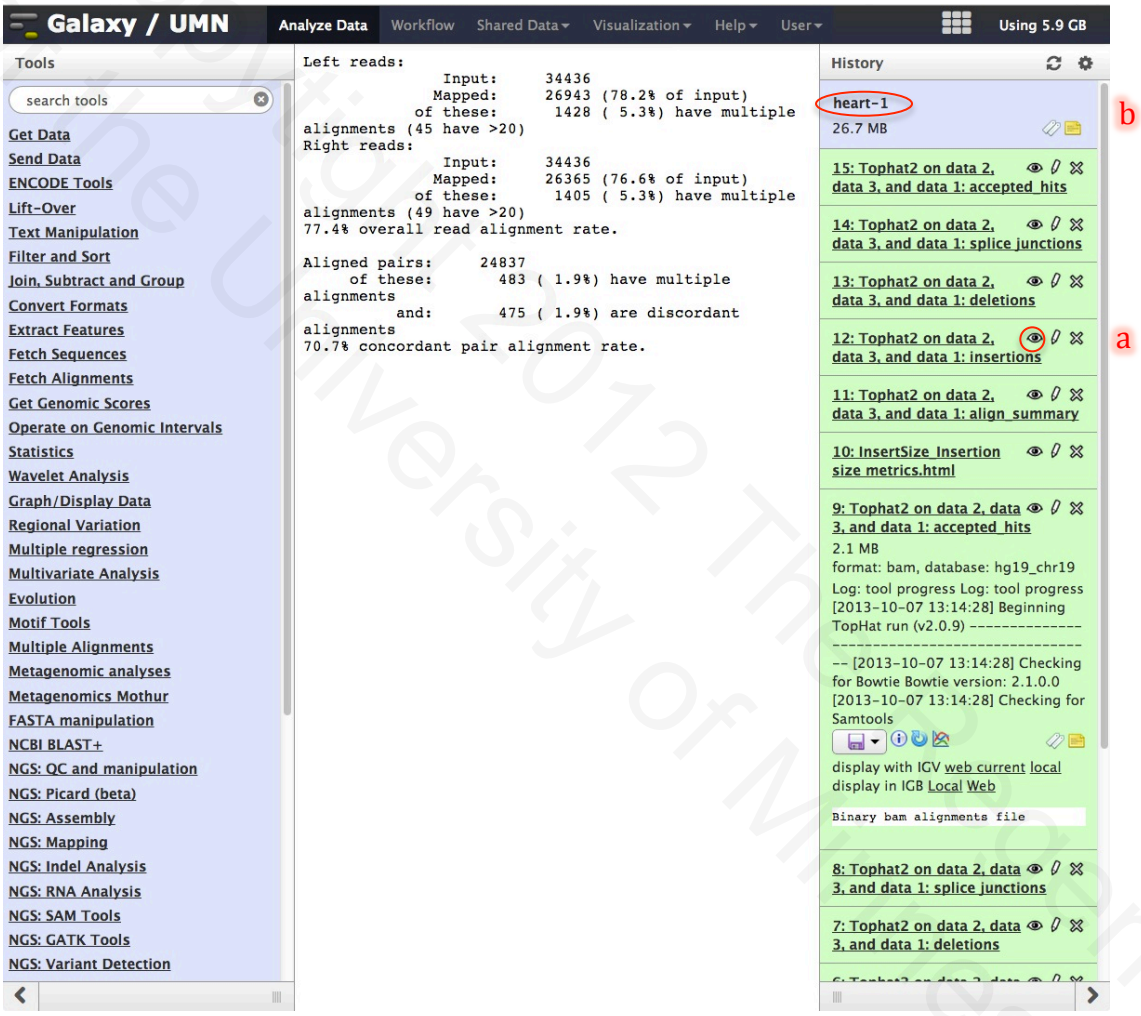


3.4 Review mapping statistics

★Mapping Statistics

a) Click on the "eye" icon next to the Tophat2 "align_summary" output file in the history pane to view the output in the central pane

b) Rename the current history: at the top of the history pane click on "Unnamed history" and rename it "heart-1". (NOTE: you must hit 'Enter' after typing the new name, rather than clicking outside the box)



4 Workflows

★Galaxy Workflows (Sect 8 page 27)
All of the steps that have been performed on the heart-1 sample need to be repeated, in separate histories, for the two other heart samples and the three skeletal samples. Galaxy workflows provide an easy method to automate an analysis pipeline. Appendix A demonstrates how to generate a workflow from your current history and use it to analyze another sample. To save time we will not work through this section in the hands-on workshop, but this section should be completed if working on a real dataset.

5 Visualizing alignments with IGV

★Visualization (Sect 5.3 page 18)
Visualizing alignments is a quick and easy way to check for major problems with the data. You may wish to verify that housekeeping genes are indeed roughly evenly covered with reads, or documented differentially-expressed genes indeed have differential coverage between samples of different groups.

★Galaxy Visualization Options (Sect 5.2 page 17)
Galaxy supports three genome browsers for visualizing data:
The Integrative Genomics Viewer (IGV) is the recommended genome browser because it is fast, powerful, and easy to use.
Trackster is a genome browser built into Galaxy. Any data file that can be viewed in Trackster will have a Trackster icon  next to it in the history pane.
The Integrated Genome Browser (IGB) is similar to IGV, but most users prefer to use IGV.

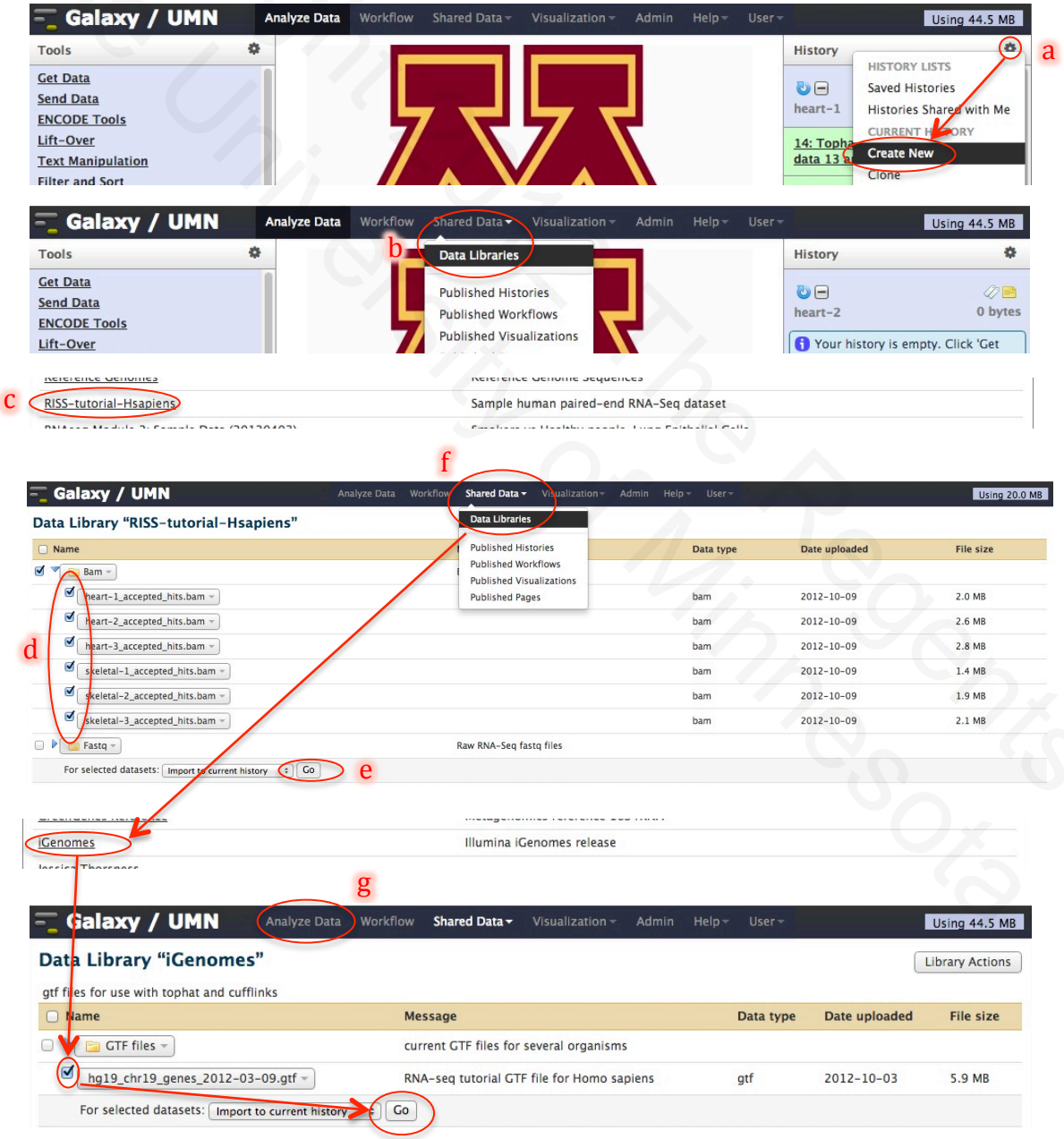
★Sample Dataset (Sect 5.1 page 16)
In this section we start with Bam alignment files that have already been generated for all six heart and skeletal samples. These Bam files were generated using the workflow previously described in this tutorial.

5.1 Load BAM alignment files and GTF into new history

★Sample Dataset

a) Create a new history by clicking on the gear icon at the top of the history window and selecting “Create New” from the drop-down menu
b) Click on “Shared Data -> Data Libraries” at the top of the window
c) Click on the “RISS-tutorial-Hsapiens” data library
d) Expand the “Bam” folder and check the box next to each bam file
e) Click “Import to current history” near the bottom of the center pane
f) Import the hg19_chr19 GTF file by clicking on “Shared Data -> Data Libraries” at the top of the screen and selecting “hg19_chr19_genes_2012-03-09.gtf” from the “iGenomes” data library
g) Return to your history by clicking on “Analyze Data” at the top of the screen

---



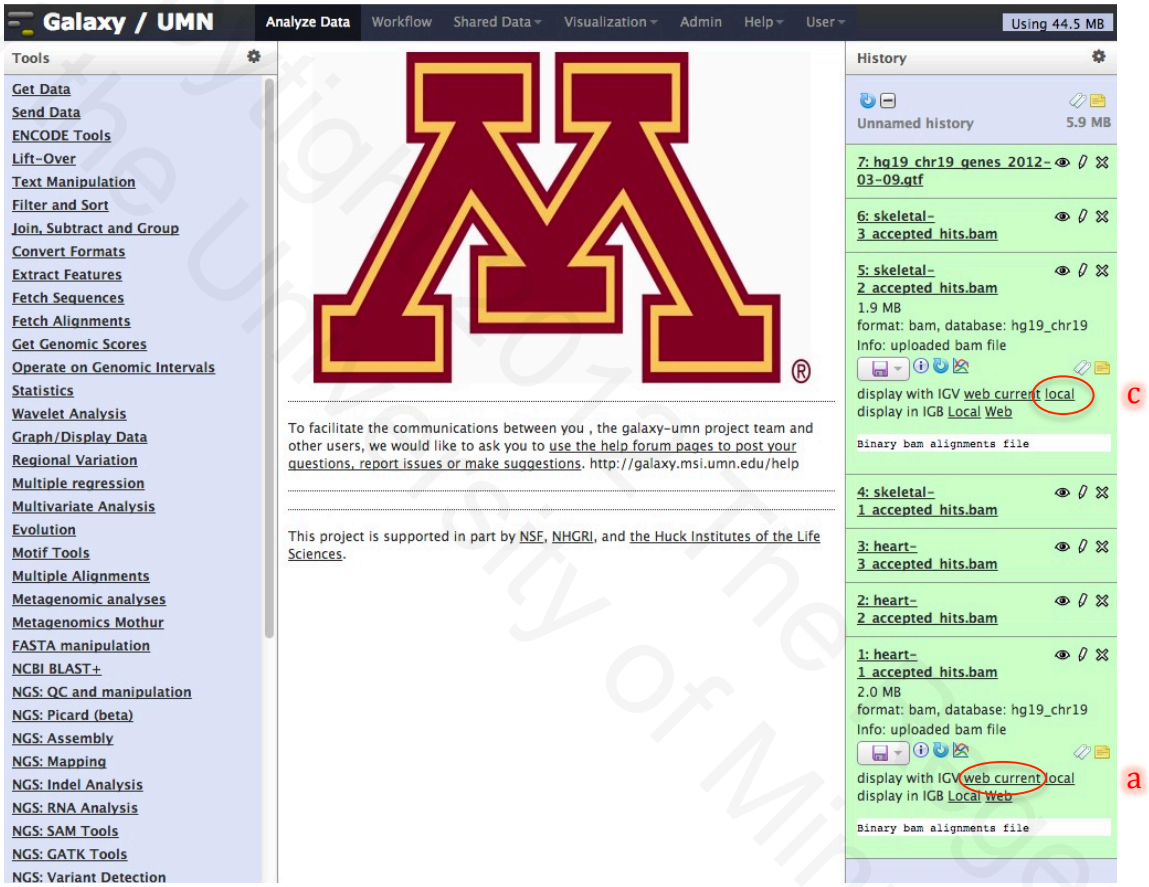
5.2 Load files into IGV

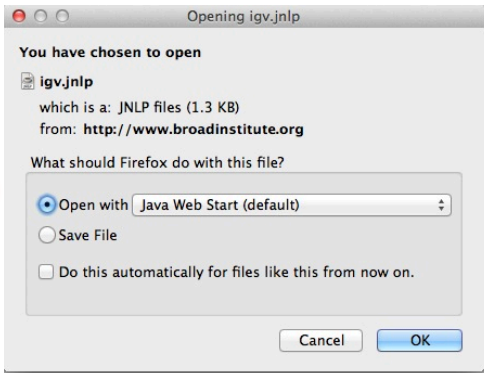
★Galaxy Visualization Options

a) Click on the "heart-1_accepted_hits.bam" file in the history pane to expand it and click on the "web current" link next to "display with IGV"

b) A file named "igv.jnlp" will be downloaded by your browser. Double click on the downloaded file to start up IGV with the heart-1.bam file loaded

c) In Galaxy click on the "skeletal-1_accepted_hits.bam" file in the history pane to expand it and click on the "local" link next to "display with IGV". The skeletal-1.bam file will load into IGV.



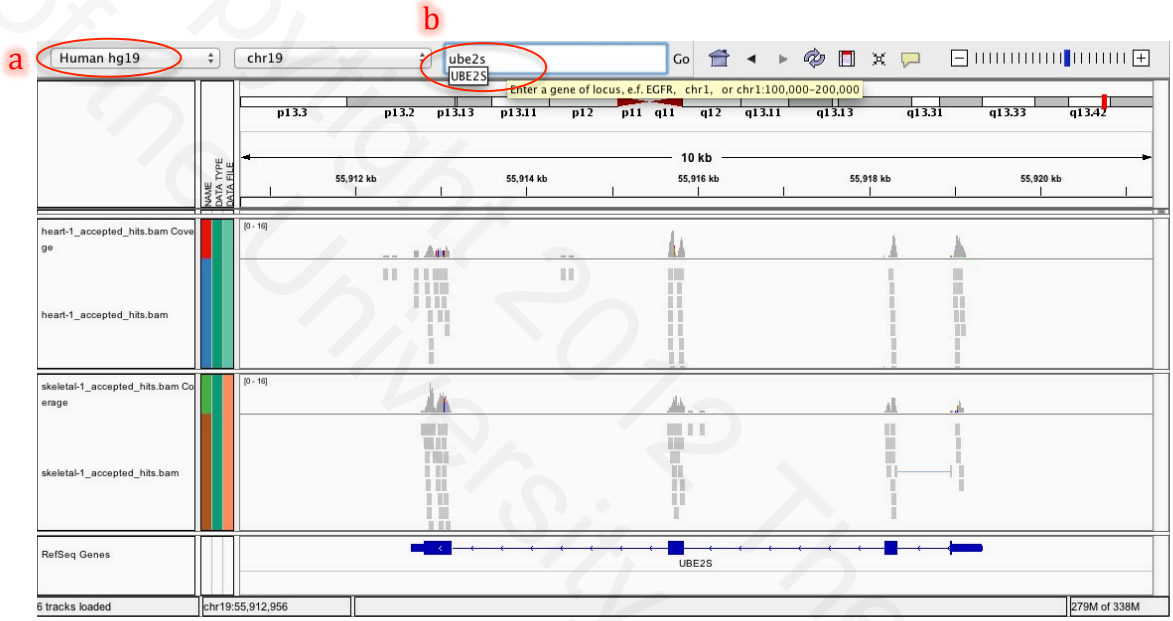


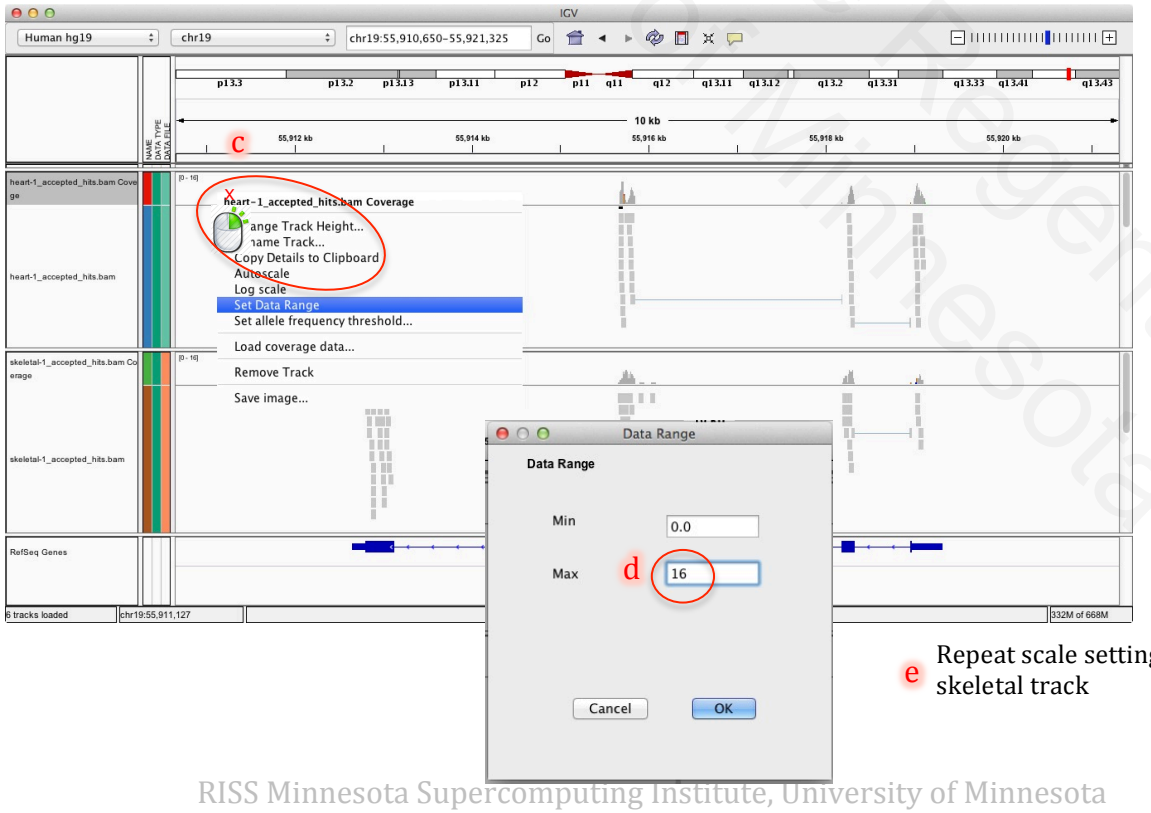
b

5.3 Look at a housekeeping gene

★Visualization

a) Verify that "Human hg19" is selected as the reference genome from the drop-down menu at the top left of the IGV window
b) Enter "ube2s" in the search box to view the reads aligning to the ubiquitin-conjugating enzyme E2S gene, which is expected to have similar express levels in both tissue types
c) Right-click on the heart coverage track and select "Set Data Range"
d) Set the "Max" value to 16
e) Repeat for the skeletal coverage track

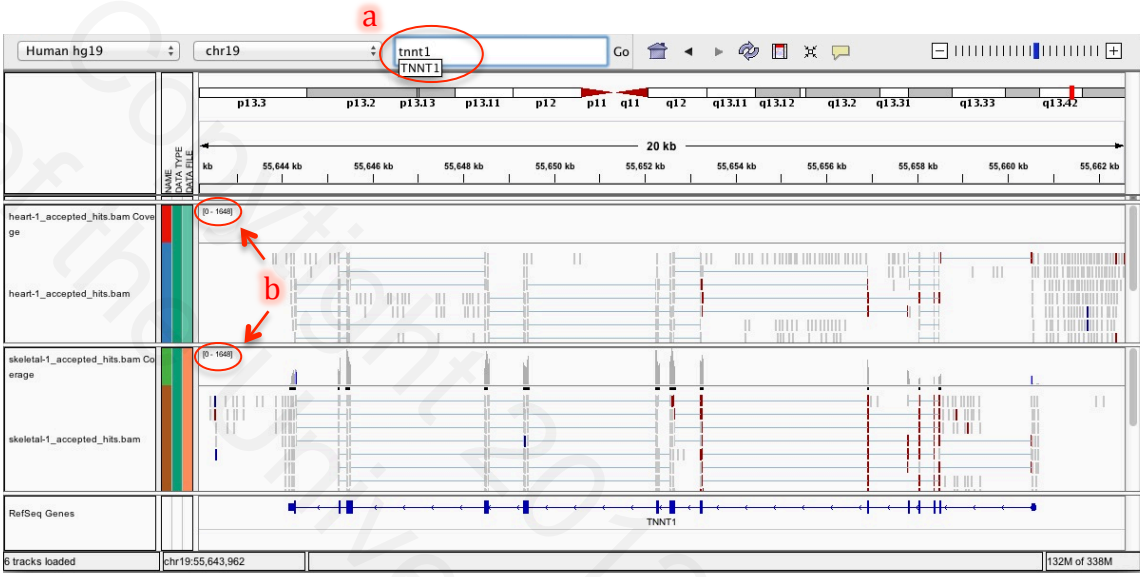




e Repeat scale setting on skeletal track

5.4 Look at a gene with differential expression

a) Enter “tnnt1” in the search box to view the reads aligning to the Troponin T, slow skeletal muscle gene, which is expected to be expressed only in skeletal muscle
b) Adjust the scale of the coverage tracks as needed (try max=1700)



6 Computing differential expression with cuffdiff

★**Cuffdiff Output** (Sect 6.2 page 22)
Cuffdiff produces many output files. In this tutorial we look at the gene differential expression testing file which shows which genes are differentially expressed. The other output files also contain important data, including the results of differential expression testing for spliced transcripts, primary transcripts, and coding sequences. See the cufflinks manual for detailed information about what information is in each file: cufflinks.cbcb.umd.edu/manual.html - cuffdiff_output

★**Differential Gene Expression** (Sect 6.2 page 22)
The gene differential expression testing output file is a tab-delimited text file with one row for each gene. Our sample dataset only covers a small portion of chr19 so most genes will have too few aligned reads for a differential expression test. These genes are indicated with "NOTEST" or "LOWDATA" in column 7.

★***De novo* gene/transcript discovery** (Sect 6.1 page 21)
The analysis pipeline used in this tutorial will quantify the expression of known genes in a reference annotation. If you are interested in discovering novel genes or spliceforms more steps need to be added to the pipeline. Refer to the Nature Protocols paper "*Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks*" for more information: www.ncbi.nlm.nih.gov/pubmed/22383036

6.1 Run cuffdiff

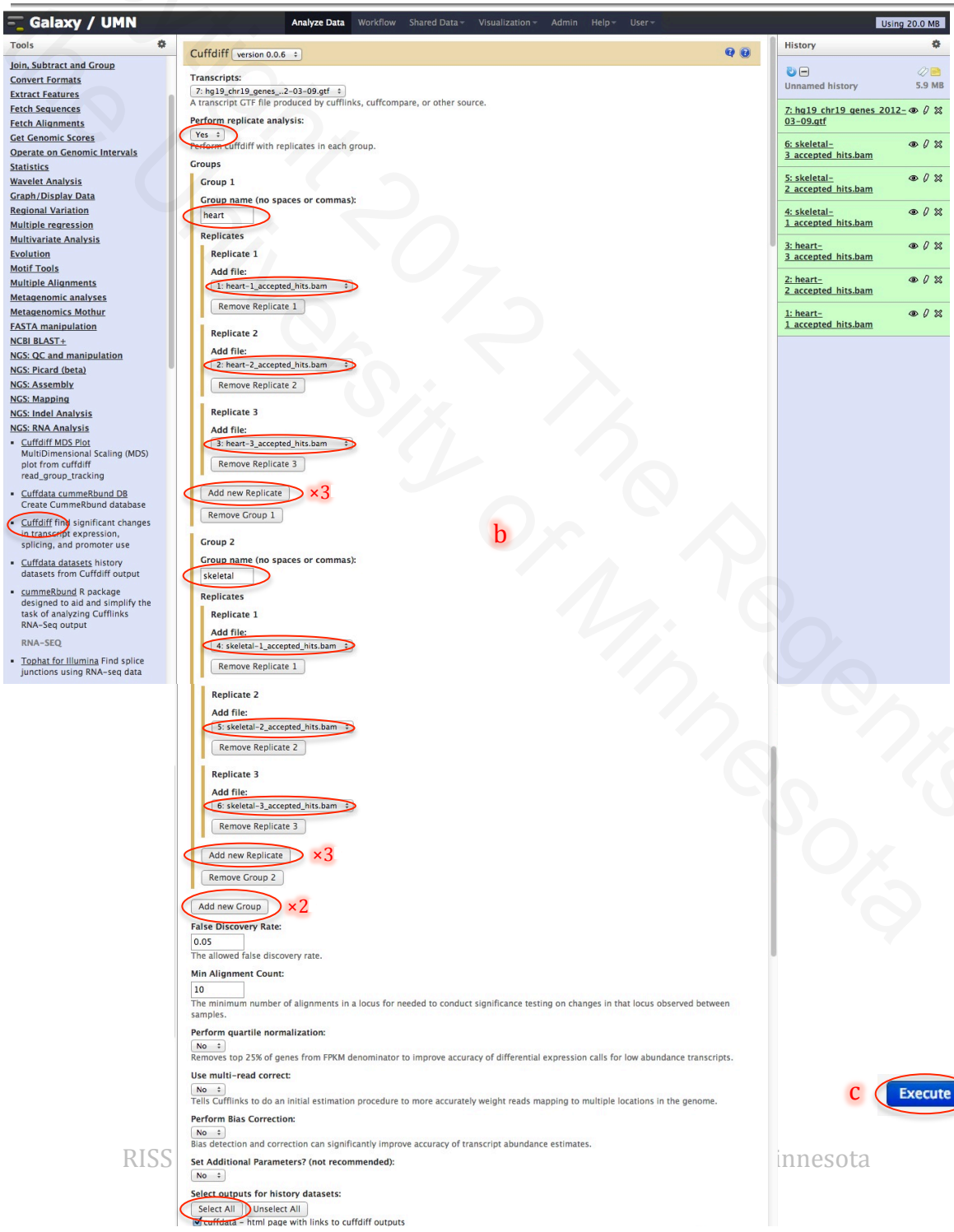
★*De novo* gene/transcript discovery

a) Load the Cuffdiff tool: "NGS: RNA Analysis -> Cuffdiff"

b) Set parameters:

- Perform replicate analysis -> Yes
- Add new Group (click twice to create two groups)
- Name Group1 "heart" and Group 2 "skeletal"
- Add new Replicate (click three times in each group)
- Set the three heart bam files as the three replicates in Group 1, and the three skeletal bam files as the three replicates in Group 2
- Select outputs for history datasets -> click "Select All"

c) Click "Execute" to submit the job



6.2 View and filter cuffdiff output

★Cufdiff Output

a) View the Cuffdiff output file “gene differential expression testing” by clicking on the “eye” icon next to the filename in the history pane

★Differential Gene Expression

b) Load the text filter tool: “Filter and Sort -> Filter”

c) Click on the output file “gene differential expression testing” to expand it in the history pane (this allows you to see the column names and numbers)

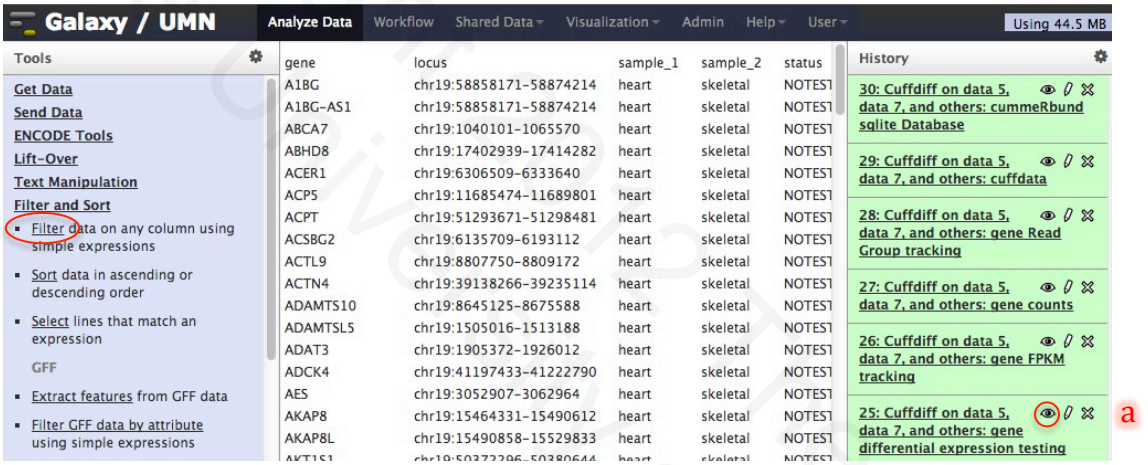
d) Set the Cuffdiff output file “gene differential expression testing” as the file to filter

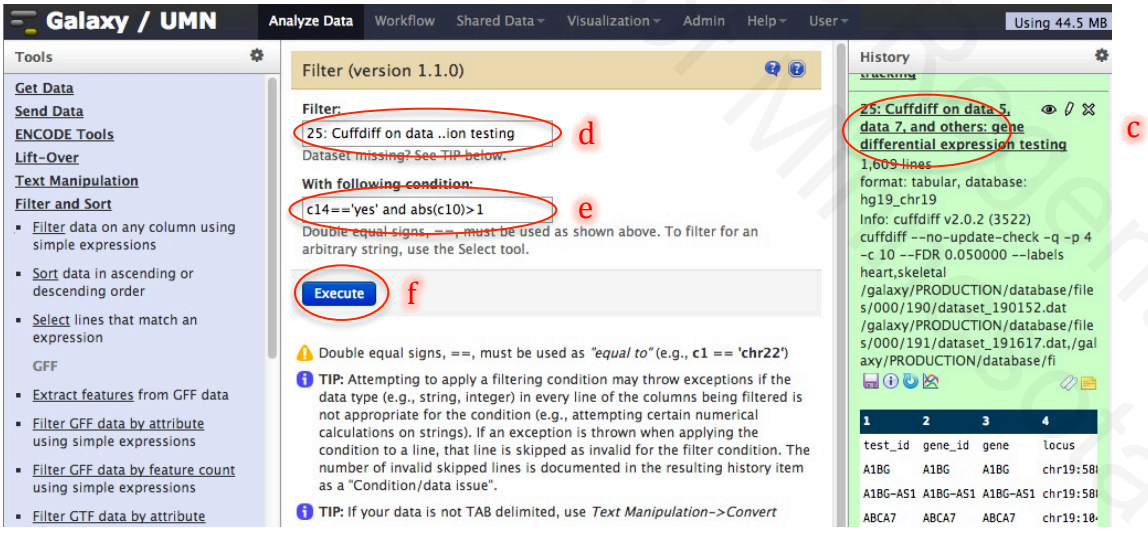
e) Filter out genes with significant change in expression with a log fold-change of at least 1 by entering “c14 == ‘yes’ and abs(c10)>1” in the “with following condition” text box

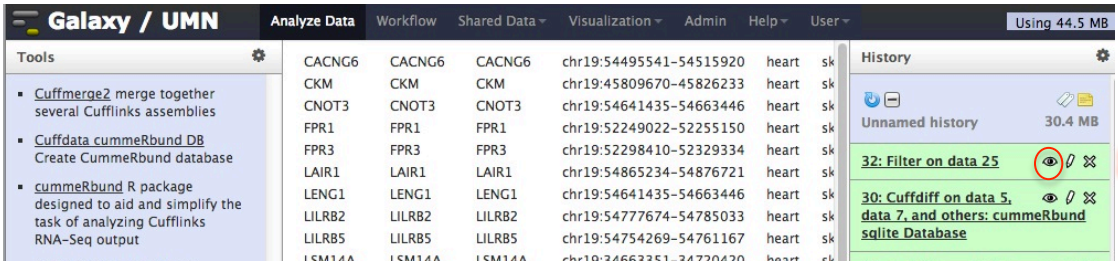
f) Click “Execute” to submit the job

g) Click on the “eye” icon next to the filter output filename to view the results in the center pane

---







7 Cuffdiff visualization with CummeRbund

★CummeRbund
CummeRbund is an easy to use R package that takes the output files from a cuffdiff run and creates a SQLite database of the results. This allows the user to explore data for genes, transcripts, transcription start sites, and CDS regions across multiple samples or conditions. CummeRbund implements numerous plotting functions for commonly used visualizations. The CummeRbund wrapper in Galaxy allows easy access to much of CummeRbund's functionality. For more details about available plots refer to the CummeRbund website: compbio.mit.edu/cummeRbund/

★Density Plots
A Kernel density plot is interpreted the same as a histogram. The density plot shows the distribution of gene expression levels across different samples. All samples should have reasonably similar distributions. A log10(FPKM) of 0 = 1 FPKM, which is very low expression.

★MDS Plots
MDS plots are similar to Principle Component Analysis (PCA) plots. They are useful for determining the major sources of variation in the dataset. Ideally samples from the same experimental group will be clustered together in the plot indicating that experimental condition is the major source of variation. Samples might also cluster by age, batch, date, technician, or other technical aspect of the experiment.

★Dendogram
A dendogram is a tree diagram showing how sample cluster by similarity. Ideally samples from the same experimental group are clustered together.

7.1 Run CummeRbund tool

★CummeRbund

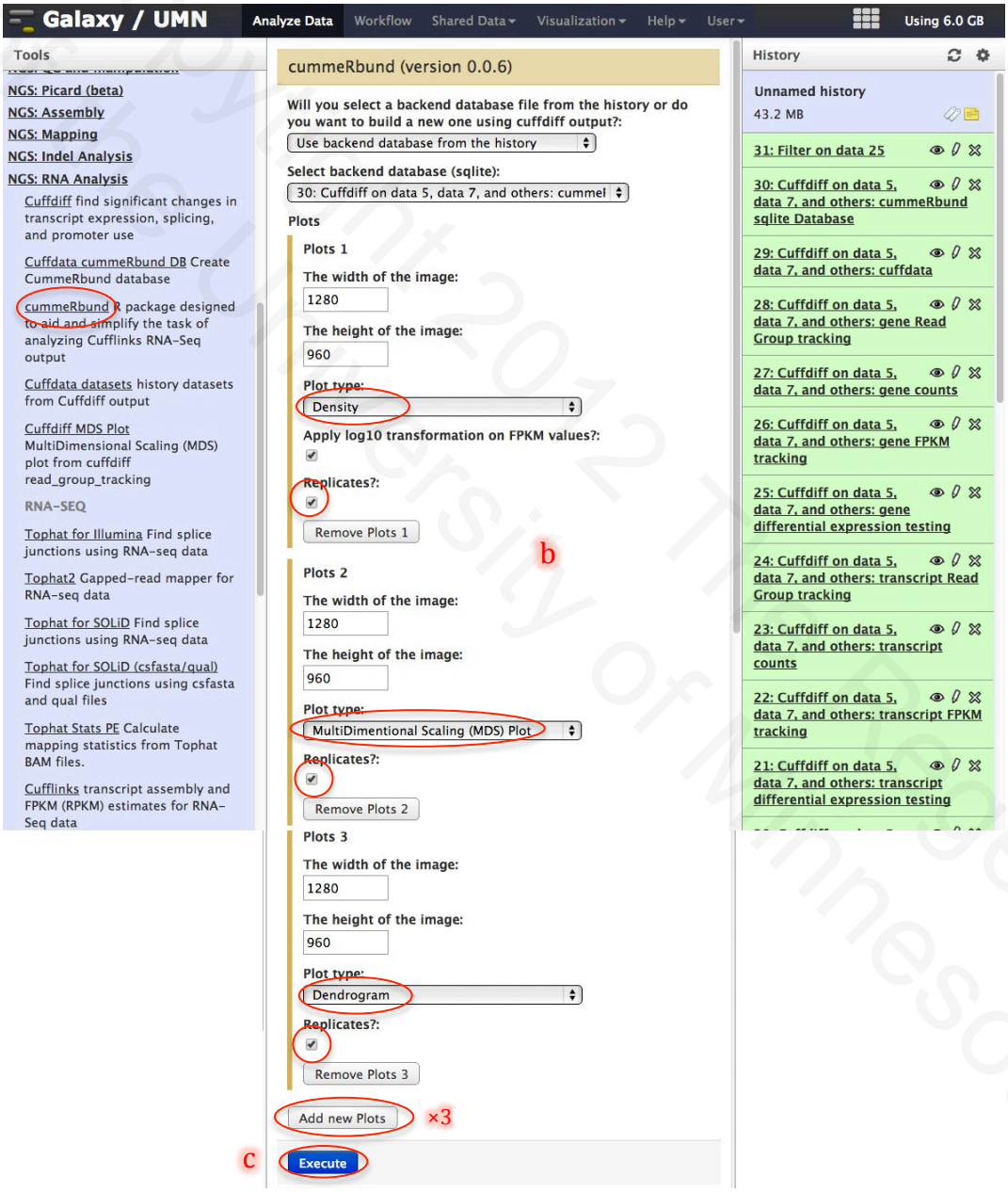
a) Load the CummeRbund tool: NGS: RNA Analysis -> cummerbund

b) Set parameters:

- Add new Plots (click three times to generate three plots)
- Plot type: Density, check the "Replicates" box
- Plot type: MultiDimensional Scaling (MDS) Plot, check the "Replicates" box
- Plot type: Dendogram, check the "Replicates" box

c) Click "Execute" to submit the job

⚠ Have patience when setting the CummeRbund parameters. After changing each setting it takes several seconds for the center pane to reload. This is common when working with large histories.



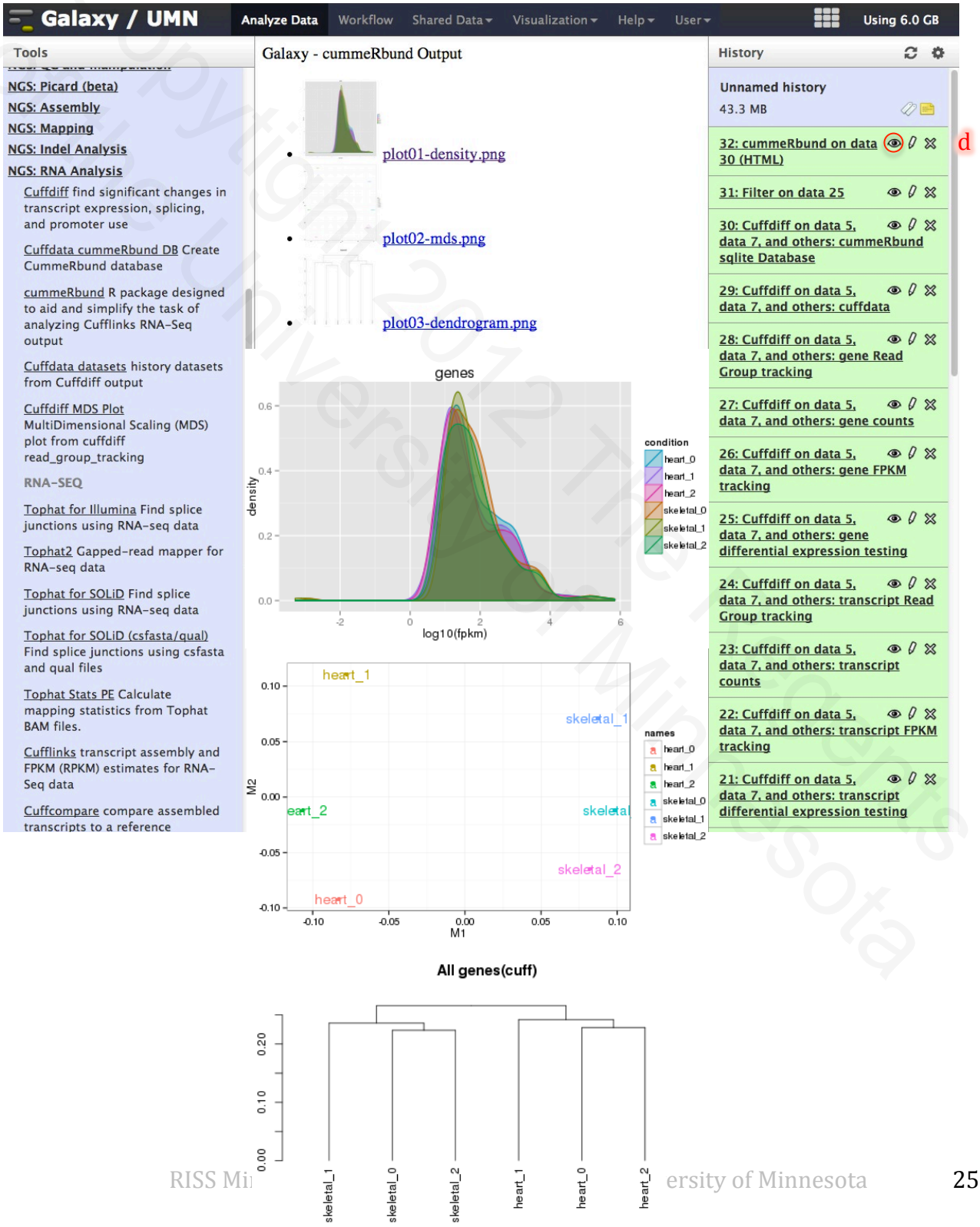
7.2 Review CummeRbund plots

★Density plots, MDS plots, and Dendograms

a) Click the "eye" icon next to the cummerbund output file to view the three plots

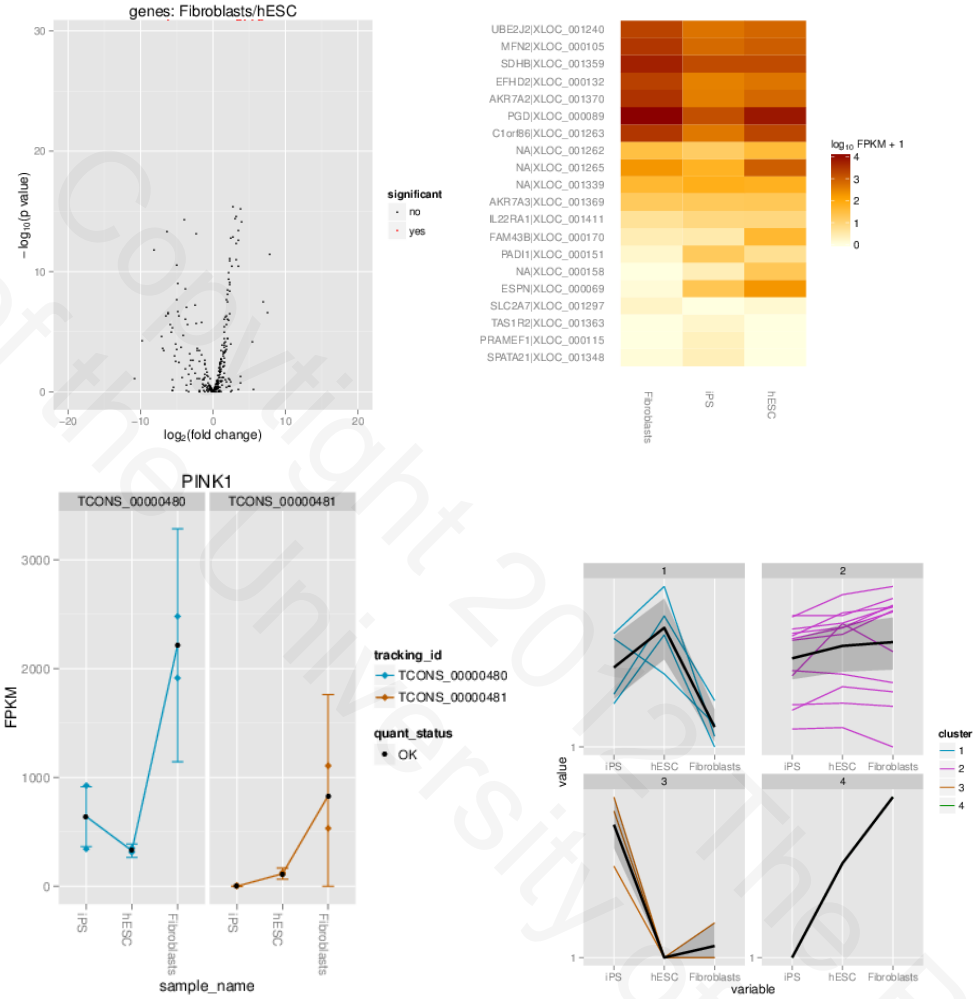
b) Verify that:

- The samples have similar density distributions
- The samples cluster by experimental condition in the MDS plot
- The sample cluster by experimental condition in the dendogram



7.3 Additional CummeRbund plots:

a) Volcano, Heatmap, Expression Plot, and Cluster.



7.4 Troubleshooting

If you experience problems using Galaxy send an email to help@msi.umn.edu with a subject beginning “RISS” and a report of the problem.

8 Appendix A: Workflows

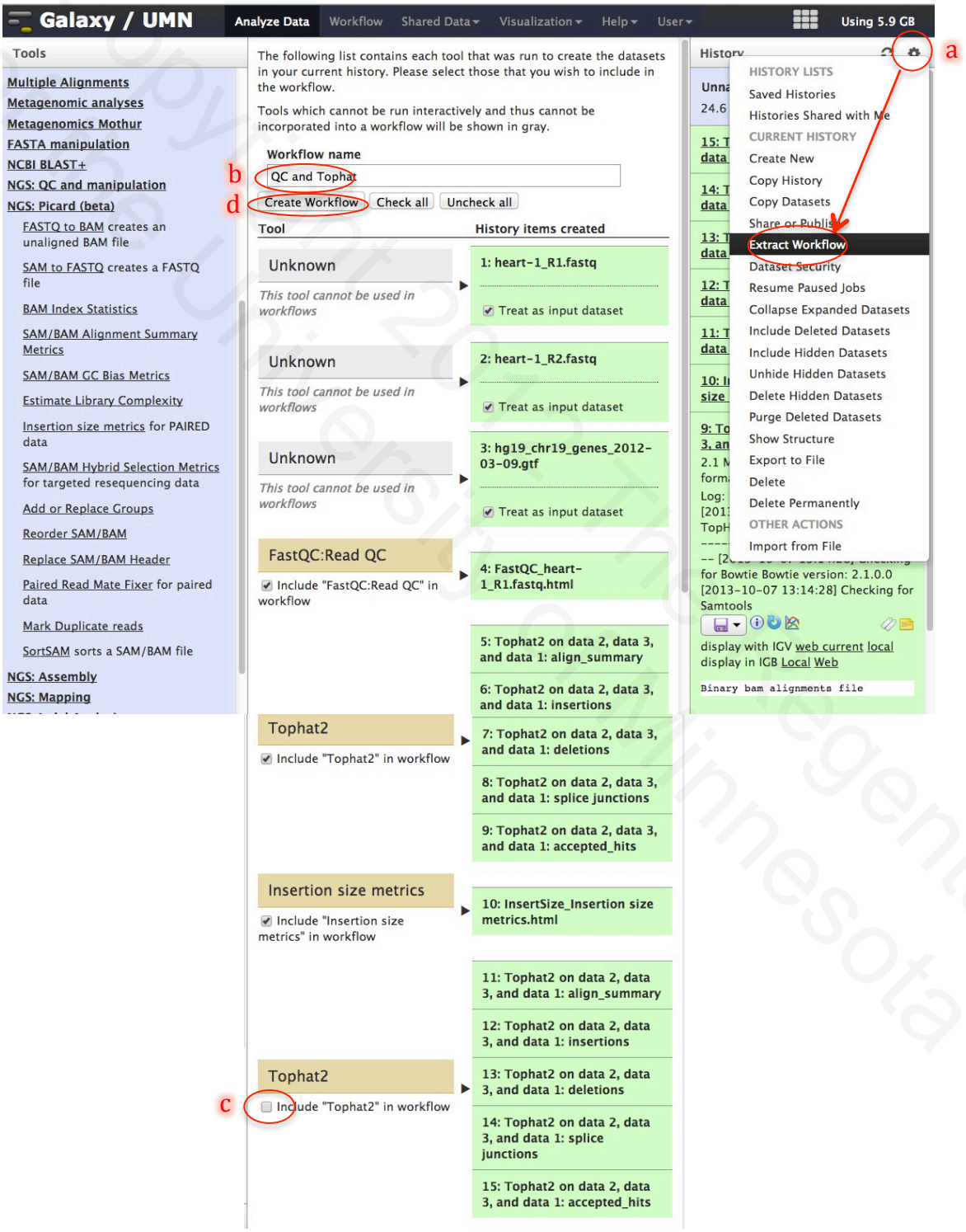
★Galaxy Workflows (Sect 8.1 page 28)
All of the steps that have been performed on the heart-1 sample need to be repeated for the two other heart samples and the three skeletal samples. Galaxy workflows provide an easy method to automate an analysis pipeline. Appendix A demonstrates how to generate a workflow from your current history and use it to analyze another sample. To save time we will not work through this section in the hands-on workshop.

★Workflow Parameters (Sect 8.2 page 30)
The workflow we set up in this section will run FastQC, Tophat2, and Insertion size metrics. Tophat2 will be run just once using the inner mate distance calculated from the first sample. Samples that were sequenced together in the same batch often have very similar average insert sizes and the same inner mate distance can be used for all samples. Check the Insertion size metrics results after running the workflow to verify that is the case.

8.1 Extract workflow from current history

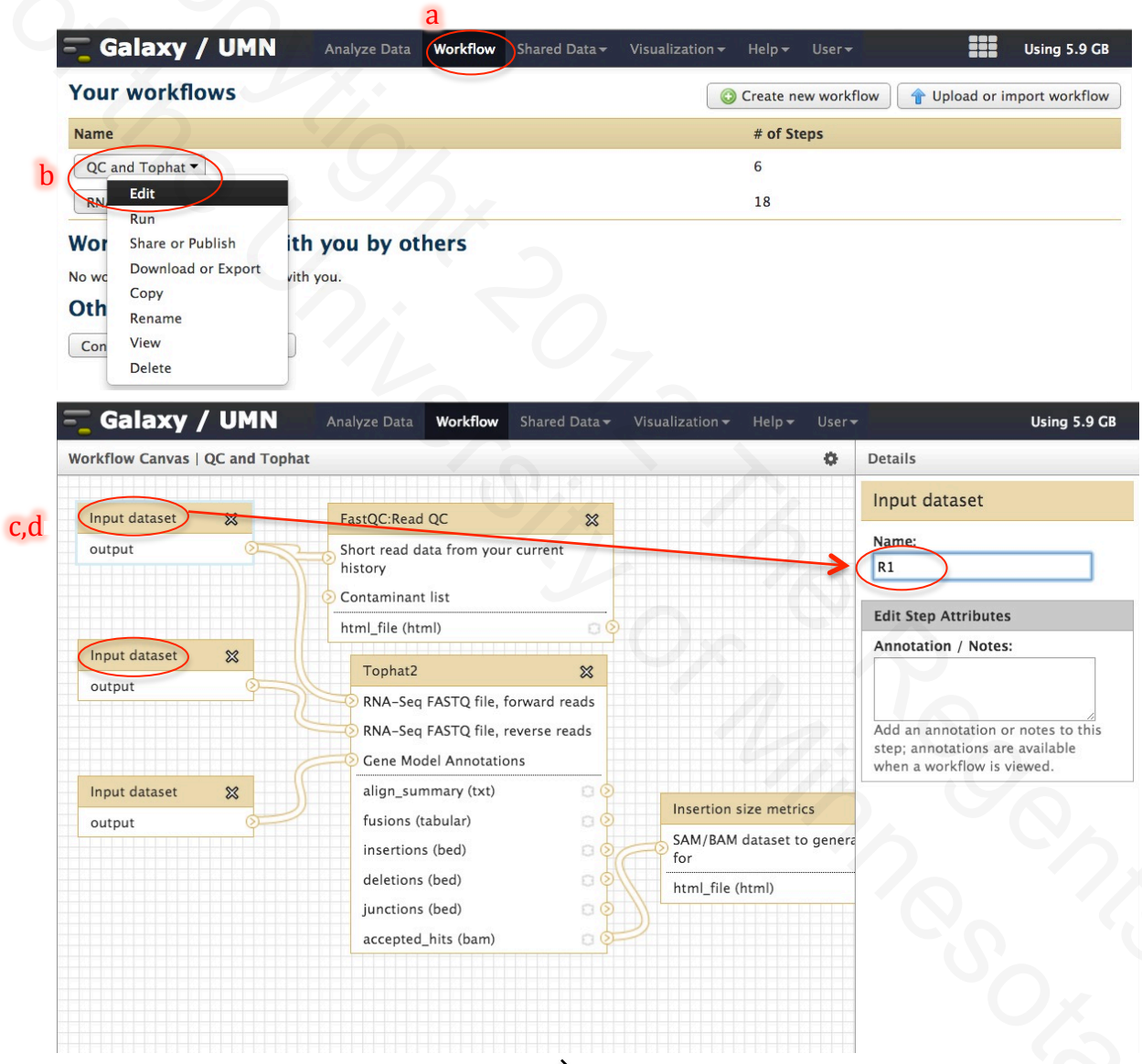
★Galaxy Workflows

a) At the top of the history pane click on the small gear icon and select “Extract Workflow” from the pop-up menu
b) In the “Workflow name” box enter “QC and Tophat
c) Uncheck the second (closest to the bottom) Tophat2 run
d) Click ”Create Workflow”



8.2 Edit the workflow

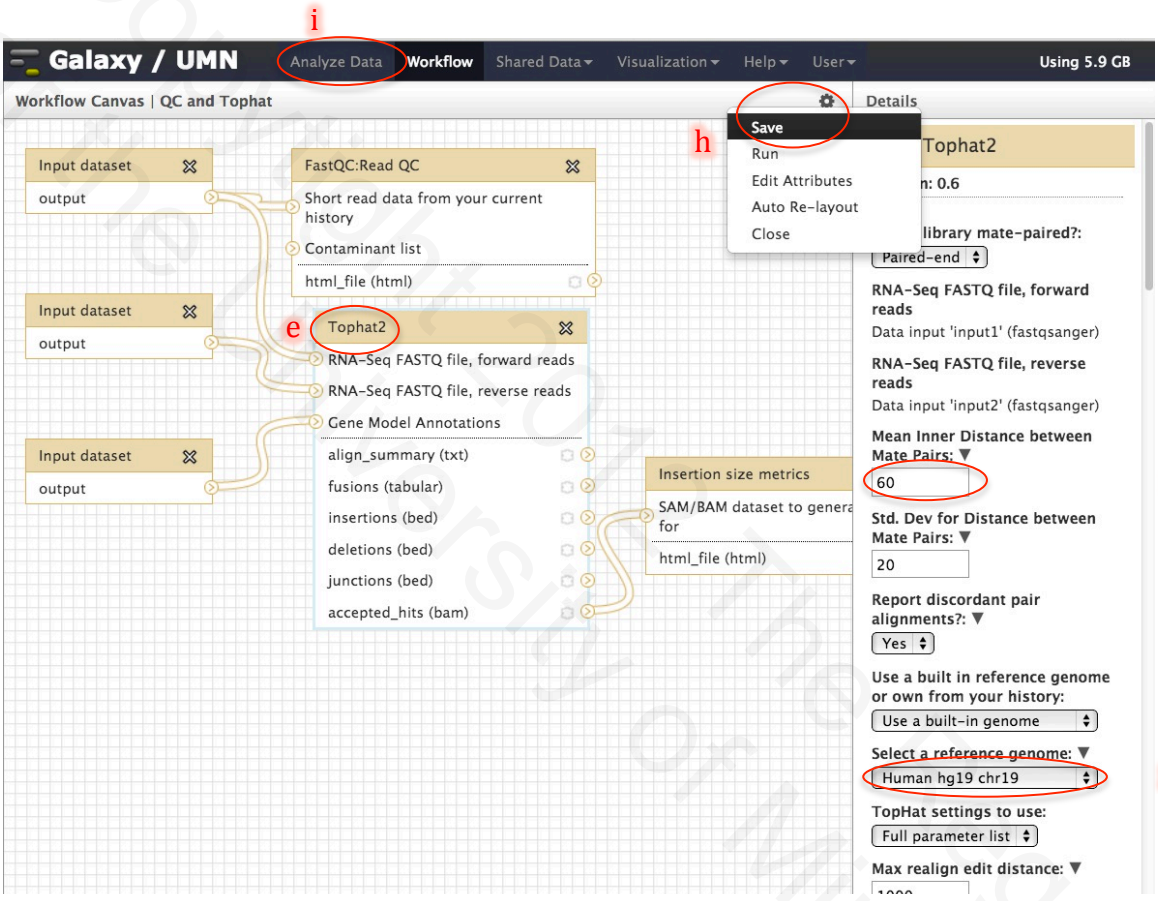
a) Click on "Workflow" at the top of the Galaxy window
b) Click on the workflow that was just created and select "Edit" from the drop-down menu
c) Move the elements of the workflow around to make it easier to see how they are connected.
d) Click on the first Input dataset box and set the Name field to 'R1'. Repeat for second input dataset ('R2').



continued on next page...

e) Click on the Tophat2 box to display the Tophat2 options in the "Details" pane on the right side.
f) Set the "Mean Inner Distance between Mate Pairs" to 60.
g) Verify the other Tophat2 parameters are set correctly.
h) Save your changes by selecting "Options -> Save" near the top of the screen
i) Return to your history by clicking on "Analyze Data" at the top of the screen

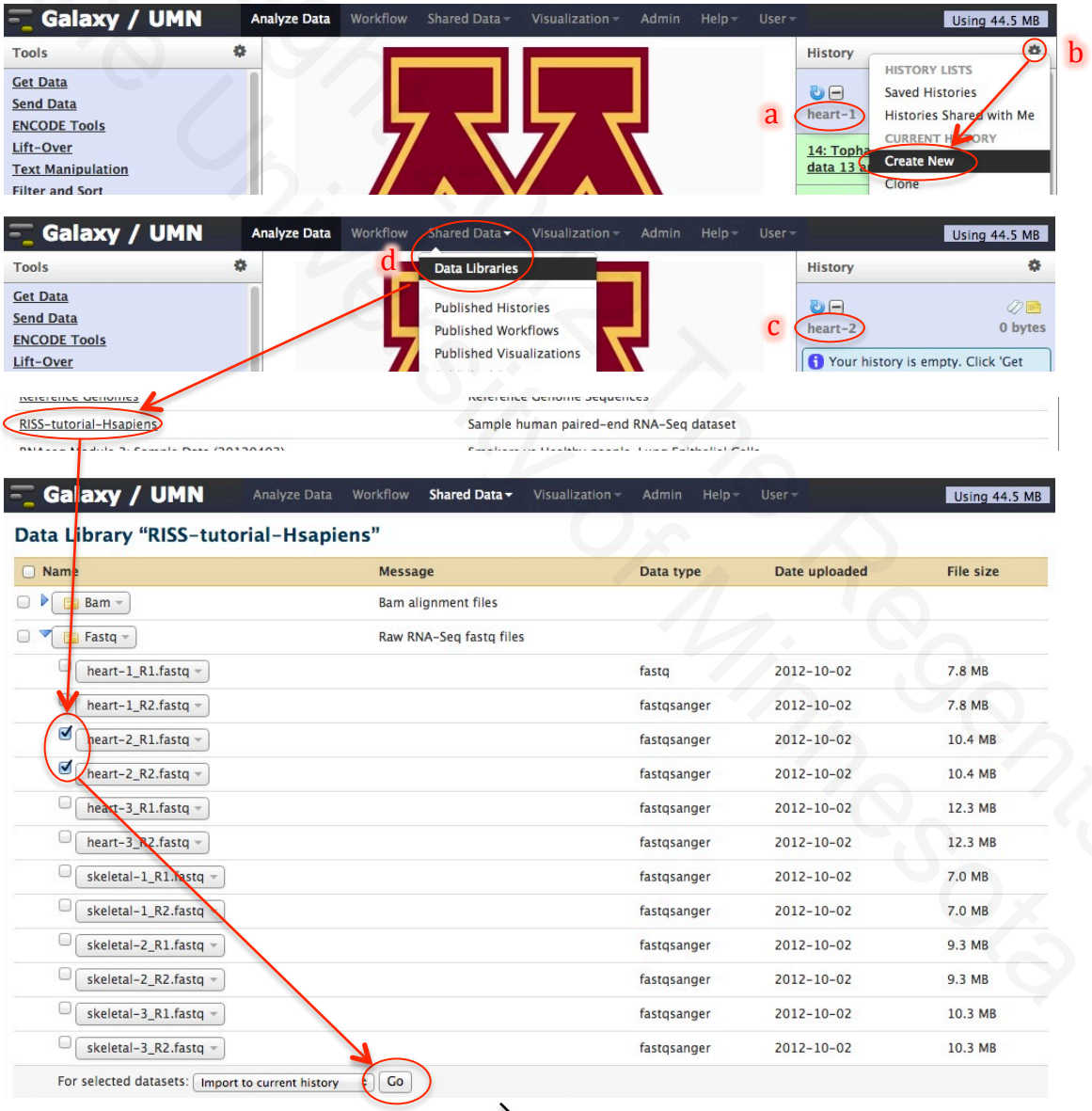
★Workflow parameters



8.3 Create new history

a) Rename the current history: at the top of the history pane click on "Unnamed history" and rename it "heart-1". (NOTE: you must hit 'Enter' after typing the new name, rather than clicking outside the box.)
b) Create a new history by clicking on the gear icon at the top of the history pane and selecting "Create New" from the pop-up menu
c) Name the new history "heart-2"
d) Import the heart-2 fastq files by clicking on "Shared Data -> Data Libraries" at the top of the screen and selecting the "heart-2_R1.fastq" and "heart-2_R2.fastq" files from the "RISS-tutorial-Hsapiens" data library

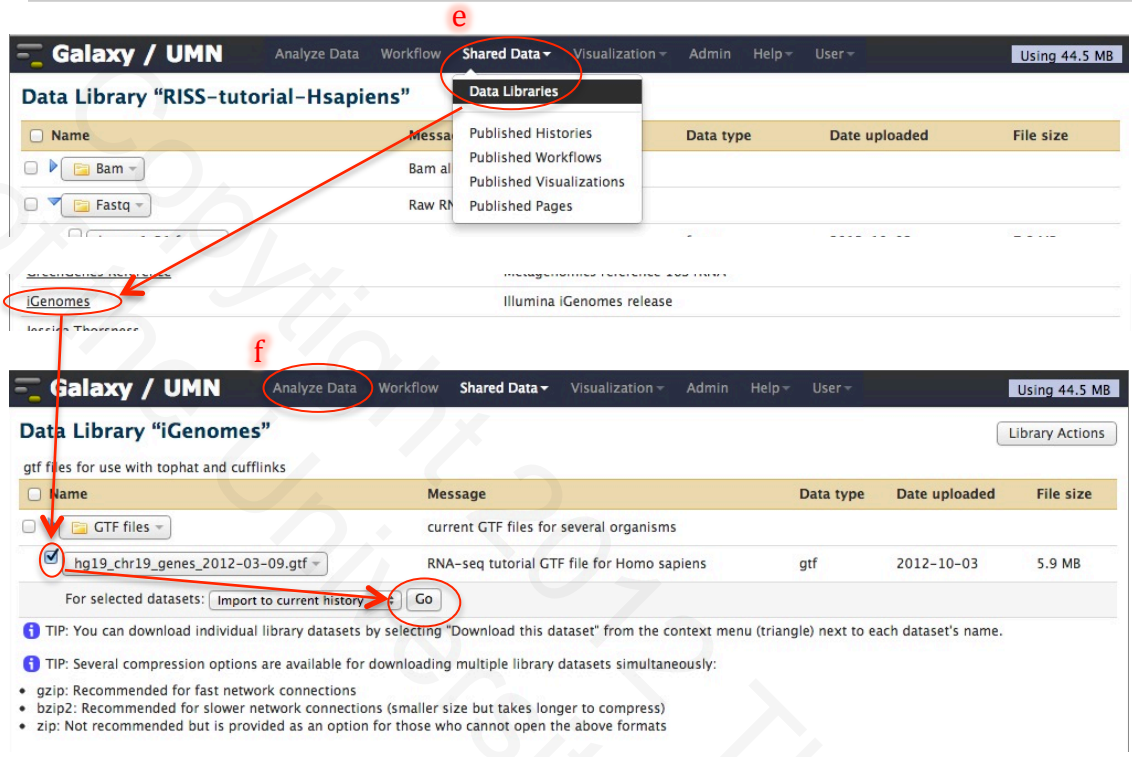
continued on next page…



continued on next page…

e) Import the hg19_chr19 GTF file by clicking on “Shared Data -> Data Libraries” at the top of the screen and selecting “hg19_chr19_genes_2012-03-09.gtf” from the “iGenomes” data library
f) Return to your history by clicking on “Analyze Data” at the top of the screen

---



8.4 Run workflow

a) Load a workflow by clicking on "Workflow" at the top of the screen
b) Click on the workflow that was just created and select "Run" from the dropdown menu
c) Select the "heart-2_R1.fastq" file in the first drop-down menu and the "heart-2_R2.fastq" file in the second drop-down menu
d) Verify the GTF file is selected in the third drop-down menu
e) Click on "Run workflow" to submit the FastQC, Tophat2, Insertion size metrics, and TophatstatsPE jobs.