

# Galaxy RNA-Seq Analysis: *H. sapiens*

---

*Tutorial*

*Research Informatics Support Systems*

*Minnesota Supercomputing Institute*

*University of Minnesota*

*Version 2*

*10/8/2013*

# Introduction

1	Introduction .....	3
1.1	<i>Scope of this tutorial</i> .....	3
1.2	<i>Reference materials</i> .....	3
1.3	<i>Outline of tutorial</i> .....	3
2	Starting Galaxy .....	4
2.1	<i>Accessing Galaxy</i> .....	5
2.2	<i>Import Fastq files for one sample into current history</i> .....	6
2.3	<i>Import the GTF file from the iGenomes data library</i> .....	7
2.4	<i>Set file attributes</i> .....	8
2.5	<i>Run FastQC</i> .....	9
3	Mapping with Tophat .....	10
3.1	<i>Initial Tophat run</i> .....	11
3.2	<i>Determine insert size</i> .....	12
3.3	<i>Rerun Tophat with correct insert size</i> .....	13
3.4	<i>Review mapping statistics</i> .....	14
4	Workflows.....	15
5	Visualizing alignments with IGV .....	15
5.1	<i>Load BAM alignment files and GTF into new history</i> .....	16
5.2	<i>Load files into IGV</i> .....	17
5.3	<i>Look at a housekeeping gene</i> .....	18
5.4	<i>Look at a gene with differential expression</i> .....	19
6	Computing differential expression with cuffdiff .....	20
6.1	<i>Run cuffdiff</i> .....	21
6.2	<i>View and filter cuffdiff output</i> .....	22
7	Cuffdiff visualization with CummeRbund .....	23
7.1	<i>Run CummeRbund tool</i> .....	24
7.2	<i>Review CummeRbund plots</i> .....	25
7.3	<i>Additional CummeRbund plots:</i> .....	26
7.4	<i>Troubleshooting</i> .....	26
8	Appendix A: Workflows .....	27
8.1	<i>Extract workflow from current history</i> .....	28
8.2	<i>Edit the workflow</i> .....	29
8.3	<i>Create new history</i> .....	31
8.4	<i>Run workflow</i> .....	33

## 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 CummeRbund 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](http://www.msi.umn.edu/content/bioinformatics-analysis)

Galaxy 101: NGS data analysis hands-on tutorial:

[www.msi.umn.edu/content/bioinformatics-analysis](http://www.msi.umn.edu/content/bioinformatics-analysis)

Tophat manual: [tophat.cbcb.umd.edu/manual.html](http://tophat.cbcb.umd.edu/manual.html)

Cufflinks manual: [cufflinks.cbcb.umd.edu/manual.html](http://cufflinks.cbcb.umd.edu/manual.html)

CummeRbund manual: [compbio.mit.edu/cummeRbund](http://compbio.mit.edu/cummeRbund)

### 1.3 Outline of tutorial

- 1 Introduction
- 2 Starting Galaxy
- 3 Mapping with Tophat
- 4 Workflows
- 5 Visualizing alignments with IGV
- 6 Computing differential expression with cuffdiff
- 7 Cuffdiff visualization with CummeRbund
- 8 Appendix A: Workflows

## 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](http://www.ensembl.org/info/data/ftp/index.html), [genome.ucsc.edu/cgi-bin/hgTables?command=start](http://genome.ucsc.edu/cgi-bin/hgTables?command=start), or NCBI. The GTF files provided in the Illumina iGenomes collection ([cufflinks.ccb.umd.edu/igenomes.html](http://cufflinks.ccb.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](http://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](http://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](http://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](http://www.msi.umn.edu/content/bioinformatics-analysis)

# Starting Galaxy

## 2.1 Accessing Galaxy

- Open a web browser and navigate to MSI Galaxy website [galaxy.msi.umn.edu](https://galaxy.msi.umn.edu)
- Log in with your MSI username and password

The image shows two screenshots of the Galaxy web application. The top screenshot, labeled 'a', shows the login page with the URL 'galaxy.msi.umn.edu' highlighted in red. The bottom screenshot, labeled 'b', shows the main Galaxy interface with three panes: 'Tools' (left), 'Center' (middle), and 'History' (right). The 'Tools' pane lists various bioinformatics tools. The 'Center' pane features a large Minnesota 'M' logo. The 'History' pane shows an empty history with a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.

a

galaxy.msi.umn.edu

Campuses : Twin Cities Crookston Duluth Morris

UNIVERSITY OF MINNESOTA  
Driven to Discover™

Minnesota Supercomputing Institute

Authorized Use Only

Username:   
Password:   
Login

Previous Page . I've forgotten my password  
I've forgotten my username

b

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Admin Help User Using 32.6 GB

Tools

search tools

Get Data  
Send Data  
ENCODE Tools  
Lift-Over  
Text Manipulation  
Filter and Sort  
Join, Subtract and Group  
Convert Formats  
Extract Features  
Fetch Sequences  
Fetch Alignments  
Get Genomic Scores  
Operate on Genomic Intervals  
Statistics  
Wavelet Analysis  
Graph/Display Data  
Regional Variation  
Multiple regression  
Multivariate Analysis  
Evolution

History

Unnamed history  
0 bytes

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

If you have questions or concerns, please e-mail [help@msi.umn.edu](mailto:help@msi.umn.edu).

Galaxy is an open, web-based platform for data intensive biomedical research. The Galaxy team is a part of BX at Penn State, and the Biology and Mathematics and Computer Science departments at Emory University. The Galaxy Project is supported in part by NHGRI, NSF, The Huck Institutes of the Life Sciences, The Institute for CyberScience at Penn State, and Emory University.

Tools pane      Center pane      History pane

# Starting Galaxy

## 2.2 Import Fastq files for one sample into current history

### ★Tutorial Dataset

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

a

b

c

d

# Starting Galaxy

## 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

**Screenshot a:** The Galaxy interface with the "Shared Data" menu open. The "Data Libraries" option is highlighted and circled in red. A dropdown menu shows "Published Histories", "Published Workflows", "Published Visualizations", and "Published Pages".

**Screenshot b:** The "Data Libraries" page. The "iGenomes" entry is circled in red. The table lists the following data libraries:

Data library name ↓	Data library description
1000Genomes	1000Genomes data from Broad
Corby_Kistler_11027	110527_SN261_0347_B81JM9ABXX/fastq_fit_syn
Fusion Sample Data	Fusion Sample Data
garbej_LVR	fahrenkr group
garbe_john_010812	test library
Genomes	Fasta files for reference genomes
GreenGenes Reference	Metagenomics reference 16S rRNA
iGenomes	Illumina iGenomes release
Metagenomics Reference	Metagenomics Reference
Mothur Examples	Mothur Examples data files

**Screenshot c:** The "Data Library "iGenomes"" page. The "hg19\_chr19\_genes\_2012-03-09.gtf" file is selected and checked (circled in red). The "Import to current history" button is highlighted and circled in red.

**Screenshot d:** The "Data Library "iGenomes"" page after import. The "Go" button is highlighted and circled in red.

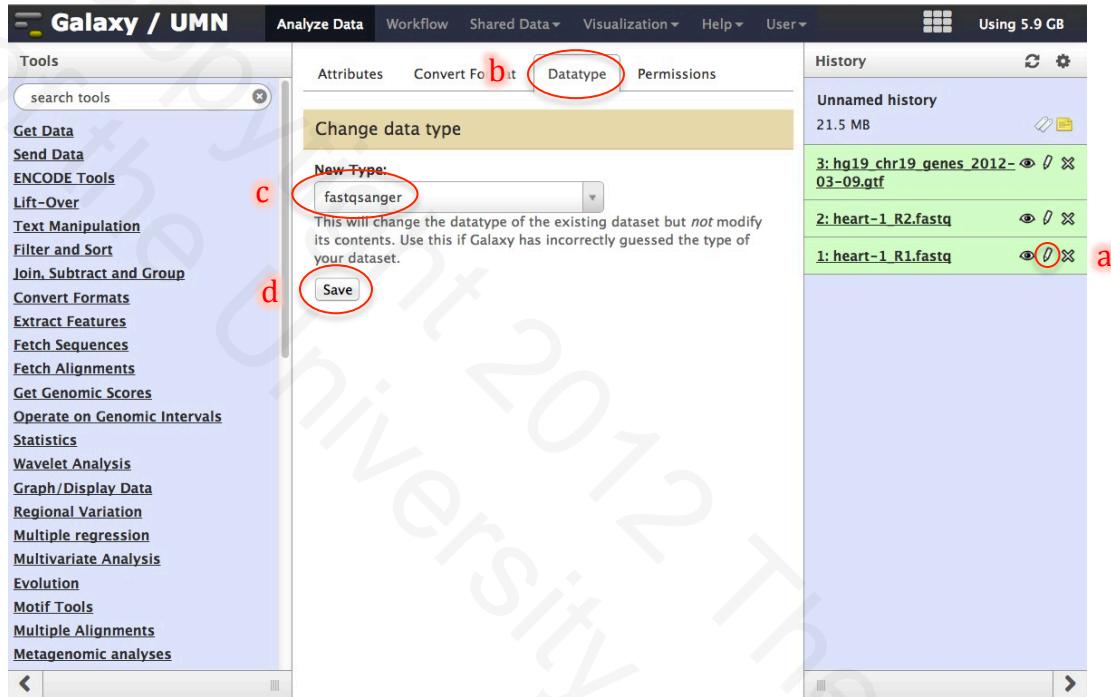
**Screenshot e:** The Galaxy interface with the "Analyze Data" tab highlighted and circled in red.

## Starting Galaxy

### 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



# Starting Galaxy

## 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" ⚠ For a real dataset you would need to repeat this step on the R2 fastq file
- d) When fastqc has finished running, click on the eye on the FastQC output file to display the file in the center pane

⚠ 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](http://www.msi.umn.edu/content/bioinformatics-analysis)

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User

History

Unnamed history  
21.5 MB

3: hg19 chr19 genes 2012-03-09.gtf

2: heart-1 R2.fastq

1: heart-1 R1.fastq

FastQC:Read QC (version 0.52)

Short read data from your current history:  
1: heart-1\_R1.fastq

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

Letters and numbers only please – other characters will be removed

Contaminant list:  
Selection is Optional

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

Execute

Purpose

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

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User

History

Unnamed history  
21.5 MB

4: FastQC\_heart-1\_R1.fastq.html

3: hg19 chr19 genes 2012-03-09.gtf

2: heart-1 R2.fastq

1: heart-1 R1.fastq

heart-1\_R1.fastq FastQC Report

FastQC Report  
Mon 7 Oct 2013  
heart-1\_R1.fastq

Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

# Mapping with Tophat

## 3 Mapping with Tophat

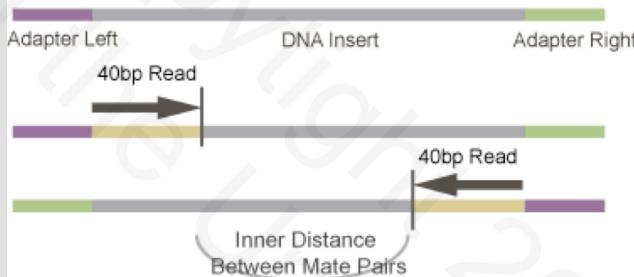
### ★Reference Genomes ([Sect 3.1 page 11](#))

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](http://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  $200 - (2 * \text{read length})$  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](#))

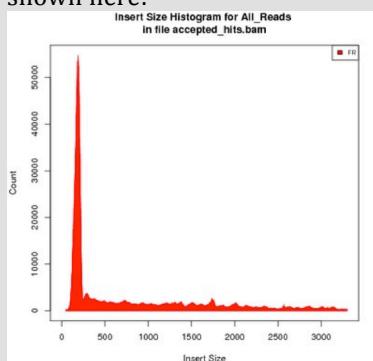
See the RNA-Seq Lecture 2 handout for more detail on setting parameters properly for other organisms: [www.msi.umn.edu/content/bioinformatics-analysis](http://www.msi.umn.edu/content/bioinformatics-analysis)

### ★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.

# Mapping with Tophat

## 3.1 Initial Tophat run

### ★Reference Genomes

### ★Mean Inner Distance – Part I

### ★Junctions

### ★Advanced Tophat Parameters

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

The screenshot shows the Galaxy web interface with the following configuration for a Tophat2 run:

- a** Tophat2 Gapped-read mapper for RNA-seq data
- b** Is this library mate-paired?: Paired-end
- c** RNA-Seq FASTQ file, forward reads: 1: heart-1\_R1.fastq
- d** RNA-Seq FASTQ file, reverse reads: 2: 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** Execute button

The History panel on the right shows the following items:

- Unnamed history
- 22.2 MB
- 4: FastQC\_heart-1\_R1.fastq.html
- 3: hg19\_chr19\_genes\_2012-03-09.gtf
- 2: heart-1\_R2.fastq
- 1: heart-1\_R1.fastq

## Mapping with Tophat

### 3.2 Determine insert size

#### ★Mean Inner Distance – Part II

#### ★Insert Size Histogram

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

The screenshot shows the Galaxy web interface with the following details:

- Tools Panel:** Shows various NGS tools including Multiple Alignments, Metagenomic analyses, Metagenomics Mothur, FASTA manipulation, NCBI BLAST+, NGS: QC and manipulation, and NGS: Picard (beta).
- Tool Configuration:**
  - Title:** Insert size metrics (version 1.56.0)
  - SAM/BAM dataset:** 9: Tophat2 on data 2, data 3, and data 1: accepted
  - Output Title:** Insert size metrics
  - Deviations:** 10.0
  - Histogram width:** 0
  - Minimum percentage:** 0.05
  - Metric Accumulation Level:** All reads (default)
- History Pane:** Displays a list of recent jobs:
  - 9: Tophat2 on data 2, data 3, and data 1: accepted hits
  - 8: Tophat2 on data 2, data 3, and data 1: splice junctions
  - 7: Tophat2 on data 2, data 3, and data 1: deletions
  - 6: Tophat2 on data 2, data 3, and data 1: insertions
  - 5: Tophat2 on data 2, data 3, and data 1: align summary
  - 4: FastQC\_heart-1\_R1.fastq.html
  - 3: hg19\_chr19\_genes\_2012-03-09.gtf
  - 2: heart-1\_R2.fastq
  - 1: heart-1\_R1.fastq

Annotations with red circles and letters:

- a**: Circles the "Insert size metrics for PAIRED data" link in the Tools panel.
- b**: Circles the "Execute" button at the bottom of the tool configuration panel.
- c**: Circles the "metrics.html" file in the History pane.

The screenshot shows the Galaxy web interface after the tool has been executed, displaying the following information:

- Tools Panel:** Same as the previous screenshot.
- Job Output:** A list of generated files:
  - 151 200
  - 152 258
  - 153 243
  - 154 211
  - 155 271
  - 156 299
  - 157 299
  - 158 316
  - 159 313
  - 160 343** (circled in red)
  - 161 277
  - 162 321
  - 163 312
  - 164 295
  - 165 316
  - 166 314
  - 167 299
  - 168 280
  - 169 267
  - 170 231
  - 171 214
  - 172 182
  - 173 186
- History Pane:** Displays a list of recent jobs, identical to the one in the first screenshot, but with the "metrics.html" file now listed under job 10.

Annotation with red circle and letter:

- d**: Circles the number 160 343 in the list of generated files.
- c**: Circles the "metrics.html" file in the History pane.

## Mapping with Tophat

### 3.3 Rerun Tophat with correct insert size

- 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
- Change the “Mean Inner Distance between Mate Pairs” to the correct value: Picard value – (2 \* read length) = 160 – (2 \* 50) = 60
- Click “Execute” to submit the job

The screenshot shows the Galaxy web interface with the following details:

- Left Panel (Tools):**
  - Multiple Alignments
  - Metagenomic analyses
  - Metagenomics Mothur
  - FASTA manipulation
  - NCBI BLAST+
  - NGS: QC and manipulation
  - NGS: Picard (beta)
  - FASTQ to BAM creates an unaligned BAM file
  - SAM to FASTQ creates a FASTQ file
  - BAM Index Statistics
  - SAM/BAM Alignment Summary Metrics
  - SAM/BAM GC Bias Metrics
  - Estimate Library Complexity
  - Insertion size metrics for PAIRED data
  - SAM/BAM Hybrid Selection Metrics for targeted resequencing data
  - Add or Replace Groups
  - Reorder SAM/BAM
  - Replace SAM/BAM Header
  - Paired Read Mate Fixer for paired data
  - Mark Duplicate reads
  - SortSAM sorts a SAM/BAM file
  - NGS: Assembly
  - NGS: Mapping
  - NGS: Indel Analysis
  - NGS: RNA Analysis
  - NGS: SAM Tools
  - NGS: GATK Tools
  - NGS: Variant Detection
- Central Panel (Tophat2 tool configuration):**
  - Tool Title:** Tophat2 (version 0.6)
  - Is this library mate-paired?**: Paired-end
  - RNA-Seq FASTQ file, forward reads:** 1: heart-1\_R1.fastq
  - Nucleotide-space:** Must have Sanger-scaled quality values with ASCII offset 33
  - RNA-Seq FASTQ file, reverse reads:** 2: heart-1\_R2.fastq
  - Nucleotide-space:** Must have Sanger-scaled quality values with ASCII offset 33
  - Mean Inner Distance between Mate Pairs:** 60 (circled with red 'b')
  - Std. Dev for Distance between Mate Pairs:** 20
  - Report discordant pair alignments?**: Yes
  - Use a built in reference genome or own from your history:** Use a built-in genome
  - Select a reference genome:** Human hg19 chr19
  - TopHat settings to use:** Full parameter list
  - Max realign edit distance:** 1000
  - Description of Max realign edit distance:** 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
- Right Panel (History):**
  - Unnamed history** (24.4 MB)
  - 10: InsertSize\_Insersion** (size metrics.html)
  - 9: Tophat2 on data 2, data 3, and data 1: accepted\_hits** (2.1 MB)
    - format: bam, database: hg19\_chr19
    - Log: tool progress Log: tool progress [2013-10-07 13:14:28] Beginning TopHat run (v2.0.9) -----
    - [2013-10-07 13:14:28] Checking for Bowtie Bowtie version: 2.1.0.0
    - [2013-10-07 13:14:28] Checking for Samtools
    - display with IGV web current local display in IGB Local Web
  - Binary bam alignments file**
  - 8: Tophat2 on data 2, data 3, and data 1: splice junctions**
  - 7: Tophat2 on data 2, data 3, and data 1: deletions**
  - 6: Tophat2 on data 2, data 3, and data 1: insertions**
  - 5: Tophat2 on data 2, data 3, and data 1: align\_summary**
  - 4: FastQC\_heart-1\_R1.fastq.html**
  - 3: hg19\_chr19\_genes\_2012-03-09.gtf**
  - 2: heart-1\_R2.fastq**

Annotations in the image:

- b**: Circles the "Mean Inner Distance between Mate Pairs" input field.
- a**: Circles the circular blue arrow icon in the history pane.
- c**: Circles the "Execute" button at the bottom of the tool configuration panel.

## Mapping with Tophat

### 3.4 Review mapping statistics

#### ★ Mapping Statistics

- 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
- 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)

The screenshot shows the Galaxy web interface with the following details:

- Tools Panel:** On the left, under the "Tools" section, various bioinformatics tools are listed, including Get Data, Send Data, ENCODE Tools, Lift-Over, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Wavelet Analysis, Graph/Display Data, Regional Variation, Multiple regression, Multivariate Analysis, Evolution, Motif Tools, Multiple Alignments, Metagenomic analyses, Metagenomics Mothur, FASTA manipulation, NCBI BLAST+, NGS: QC and manipulation, NGS: Picard (beta), NGS: Assembly, NGS: Mapping, NGS: Indel Analysis, NGS: RNA Analysis, NGS: SAM Tools, NGS: GATK Tools, and NGS: Variant Detection.
- Central Panel:** Displays the results of the "Align Data" tool. It shows statistics for Left reads and Right reads, including input counts (34436, 34436), mapped counts (26943, 26365), and percentages (78.2%, 76.6%). It also provides information on multiple alignments and discordant pairs.
- History Panel:** On the right, the history pane lists several entries:
  - 15: Tophat2 on data 2, data 3, and data 1: accepted\_hits (26.7 MB)
  - 14: Tophat2 on data 2, data 3, and data 1: splice junctions
  - 13: Tophat2 on data 2, data 3, and data 1: deletions
  - 12: Tophat2 on data 2, data 3, and data 1: insertions (circled with red 'a')
  - 11: Tophat2 on data 2, data 3, and data 1: align\_summary
  - 10: InsertSize\_Insertion size metrics.html
  - 9: Tophat2 on data 2, data 3, and data 1: accepted\_hits (2.1 MB)  
format: bam, database: hg19\_chr19  
Log: tool progress Log: tool progress [2013-10-07 13:14:28] Beginning TopHat run (v2.0.9)  
-- [2013-10-07 13:14:28] Checking for Bowtie Bowtie version: 2.1.0.0  
[2013-10-07 13:14:28] Checking for Samtools  
display with IGV web current local display in IGB Local Web
  - 8: Tophat2 on data 2, data 3, and data 1: splice junctions
  - 7: Tophat2 on data 2, data 3, and data 1: deletions
  - 6: Tophat2 on data 2, data 3, and data 1: accepted\_hits

Red circles labeled 'a' and 'b' highlight specific entries in the history pane: entry 12 (Tophat2 on data 2, data 3, and data 1: insertions) and entry 15 (Tophat2 on data 2, data 3, and data 1: accepted\_hits).

## 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.

# Visualizing alignments with IGV

## 5.1 Load BAM alignment files and GTF into new history

### ★Sample Dataset

- 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
- Click on “Shared Data -> Data Libraries” at the top of the window
- Click on the “RISS-tutorial-Hsapiens” data library
- Expand the “Bam” folder and check the box next to each bam file
- Click “Import to current history” near the bottom of the center pane
- 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
- Return to your history by clicking on “Analyze Data” at the top of the screen

The figure consists of five screenshots of the Galaxy interface, labeled a through e, illustrating the steps to load BAM and GTF files into a new history.

- Screenshot a:** Shows the Galaxy header with "Galaxy / UMN". The "History" dropdown menu is open, showing "CURRENT HISTORY" with "Create New" circled in red. A red arrow points from the "Create New" option to the "Data Libraries" button in the next screenshot.
- Screenshot b:** Shows the Galaxy header with "Galaxy / UMN". The "Shared Data" dropdown menu is open, and "Data Libraries" is highlighted and circled in red. A red arrow points from this menu to the "Data Library" section in the next screenshot.
- Screenshot c:** Shows the Galaxy header with "Galaxy / UMN". The "Data Library" section shows "RISS-tutorial-Hsapiens" circled in red. A red arrow points from this library to the "Bam" folder in the next screenshot.
- Screenshot d:** Shows the Galaxy header with "Galaxy / UMN". The "Data Library" section shows a "Bam" folder expanded, with several BAM files checked (heart-1\_accepted\_hits.bam, heart-2\_accepted\_hits.bam, heart-3\_accepted\_hits.bam, skeletal-1\_accepted\_hits.bam, skeletal-2\_accepted\_hits.bam, skeletal-3\_accepted\_hits.bam). A red circle labeled "d" is around the checked boxes. A red arrow points from this screenshot to the "Import to current history" button in the next screenshot.
- Screenshot e:** Shows the Galaxy header with "Galaxy / UMN". The "Data Library" section shows the "iGenomes" library selected, indicated by a red circle labeled "g". Below it, the "GTF files" section shows "hg19\_chr19\_genes\_2012-03-09.gtf" selected, indicated by a red circle labeled "f". A red arrow points from the "Import to current history" button in screenshot d to the "Go" button in this screenshot.

## Visualizing alignments with IGV

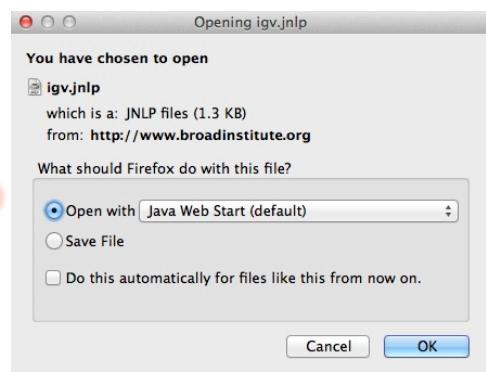
### 5.2 Load files into IGV

#### ★Galaxy Visualization Options

- 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”
- 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
- 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.

The screenshot shows the Galaxy web interface with the University of Minnesota logo in the center. The history pane on the right lists several BAM files:

- Unnamed history (5.9 MB)
- 7: hg19\_chr19\_genes\_2012-03-09.gtf
- 6: skeletal-3\_accepted\_hits.bam
- 5: skeletal-2\_accepted\_hits.bam (1.9 MB)  
format: bam, database: hg19\_chr19  
Info: uploaded bam file  
display with IGV **web current local**  
display in IGB Local Web
- Binary bam alignments file
- 4: skeletal-1\_accepted\_hits.bam
- 3: heart-3\_accepted\_hits.bam
- 2: heart-2\_accepted\_hits.bam
- 1: heart-1\_accepted\_hits.bam (2.0 MB)  
format: bam, database: hg19\_chr19  
Info: uploaded bam file  
display with IGV **web current local**  
display in IGB Local Web
- Binary bam alignments file

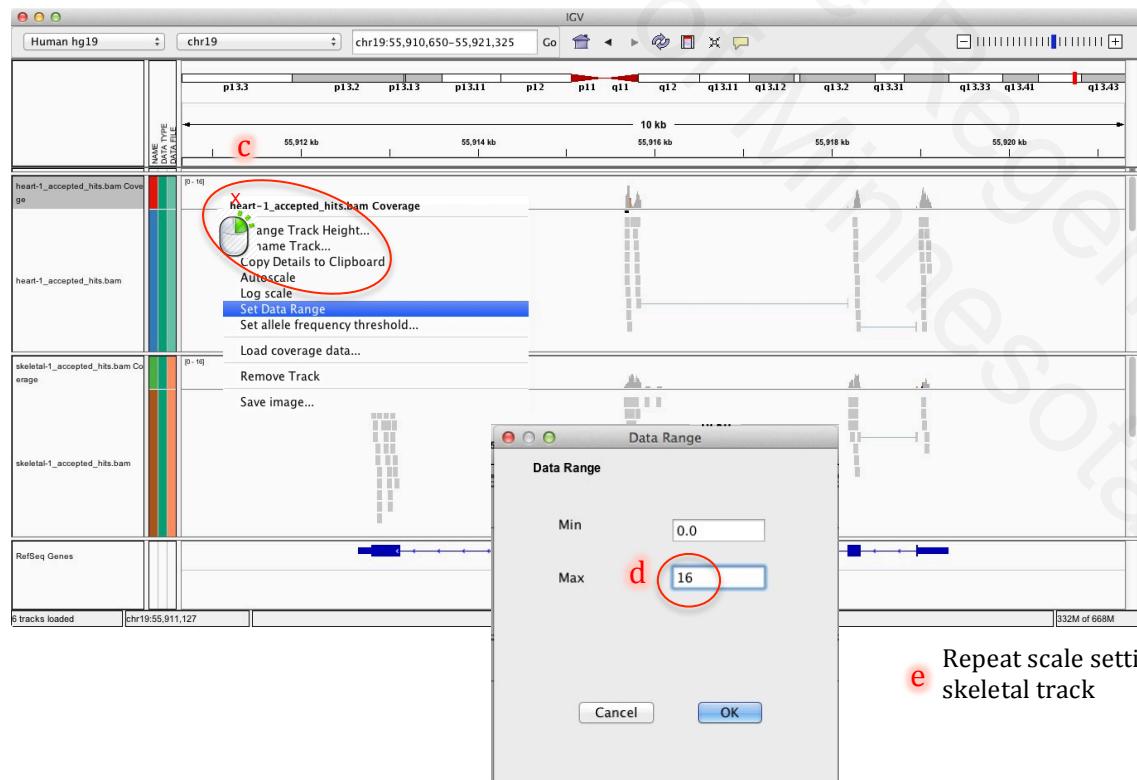
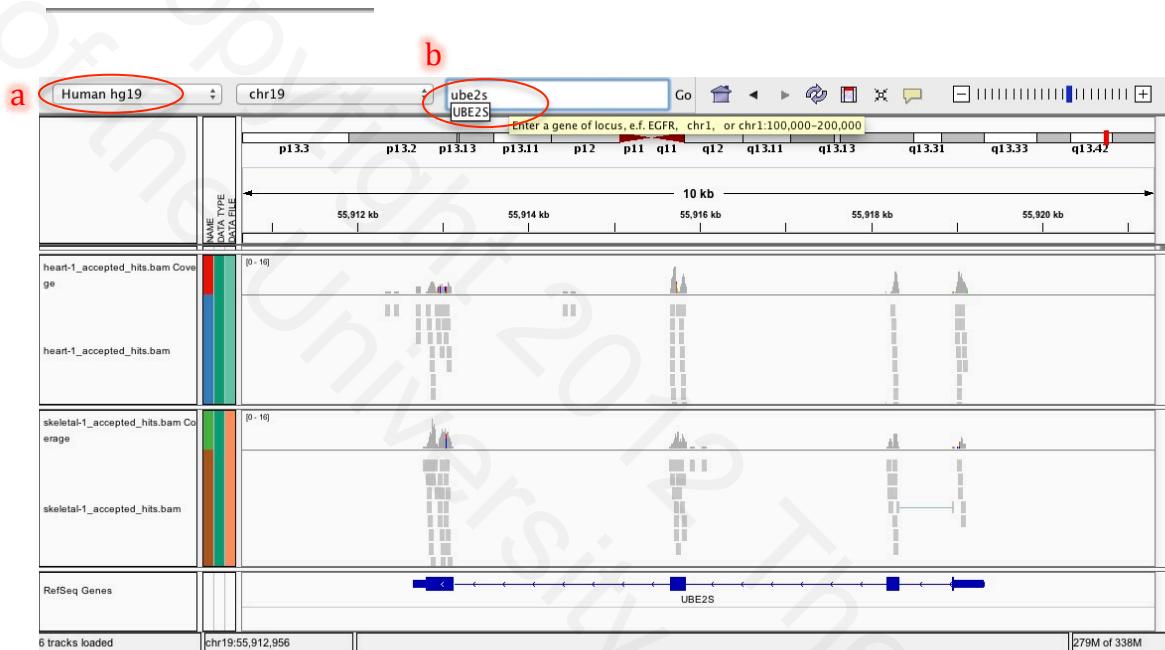


## Visualizing alignments with IGV

### 5.3 Look at a housekeeping gene

#### ★Visualization

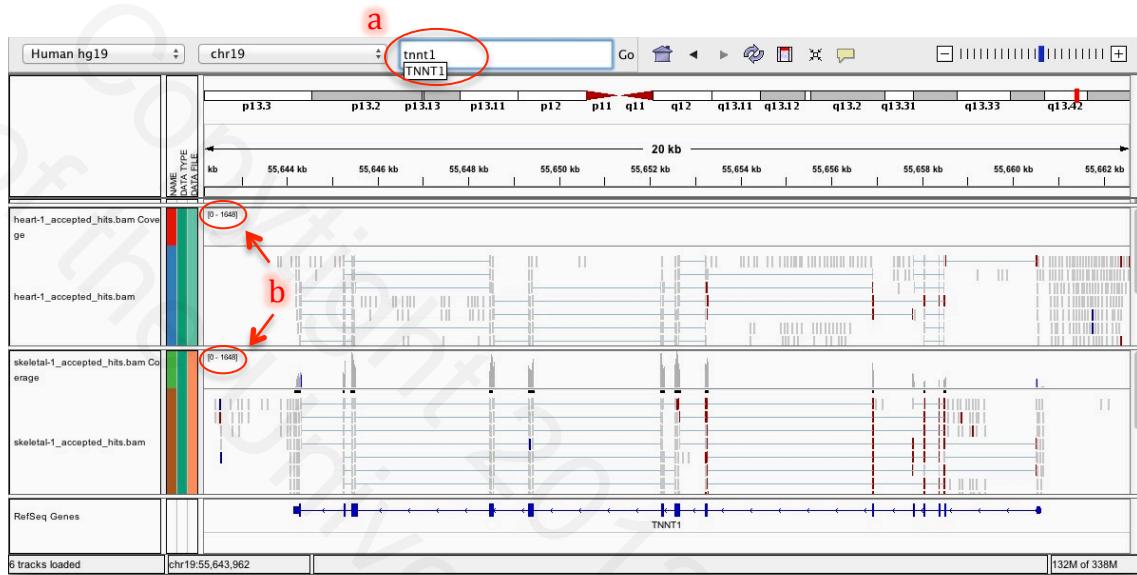
- Verify that "Human hg19" is selected as the reference genome from the drop-down menu at the top left of the IGV window
- Enter "ube2s" in the search box to view the reads aligning to the ubiquitin-conjugating enzyme E2S gene, which is expected to have similar express levels in both tissue types
- Right-click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



## Visualizing alignments with IGV

### 5.4 Look at a gene with differential expression

- 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
- 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](http://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](http://www.ncbi.nlm.nih.gov/pubmed/22383036)

# Computing differential expression with cuffdiff

## 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

The screenshot shows the Galaxy web interface with the Cuffdiff tool selected. The main panel displays the tool configuration with various parameters set. Red circles labeled 'a', 'b', and 'c' point to specific areas of interest:

- Group 1:** The 'Group name' field is circled in red, containing the value 'heart'. Below it, the 'Replicates' section shows 'Replicate 1' with an 'Add file:' dropdown containing '1: heart-1\_accepted\_hits.bam'. A red circle labeled 'a' is around the 'Add file:' dropdown.
- Group 2:** The 'Group name' field is circled in red, containing the value 'skeletal'. Below it, the 'Replicates' section shows 'Replicate 1' with an 'Add file:' dropdown containing '4: skeletal-1\_accepted\_hits.bam'. A red circle labeled 'b' is around the 'Add file:' dropdown.
- Execute:** At the bottom right of the tool configuration panel, there is a large blue 'Execute' button with a red circle labeled 'c' around it. This button is used to submit the job.

The History panel on the right lists the generated files from the job execution:

- Unnamed history (5.9 MB)
- 7: hg19\_chr19\_genes\_2012-03-09.gtf
- 6: skeletal-3\_accepted\_hits.bam
- 5: skeletal-2\_accepted\_hits.bam
- 4: skeletal-1\_accepted\_hits.bam
- 3: heart-3\_accepted\_hits.bam
- 2: heart-2\_accepted\_hits.bam
- 1: heart-1\_accepted\_hits.bam

## Computing differential expression with cuffdiff

### 6.2 View and filter cuffdiff output

#### ★Cuffdiff 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

gene	locus	sample_1	sample_2	status
A1BG	chr19:58858171-58874214	heart	skeletal	NOTEST
A1BG-AS1	chr19:58858171-58874214	heart	skeletal	NOTEST
ABCA7	chr19:1040101-1065570	heart	skeletal	NOTEST
ABHD8	chr19:17402939-17414282	heart	skeletal	NOTEST
ACER1	chr19:6306509-6333640	heart	skeletal	NOTEST
ACP5	chr19:11685474-11689801	heart	skeletal	NOTEST
ACPT	chr19:51293671-51298481	heart	skeletal	NOTEST
ACSBG2	chr19:6135709-6193112	heart	skeletal	NOTEST
ACTL9	chr19:8807750-8809172	heart	skeletal	NOTEST
ACTN4	chr19:39138266-39235114	heart	skeletal	NOTEST
ADAMTS10	chr19:8645125-8675588	heart	skeletal	NOTEST
ADAMTS15	chr19:1505016-1513188	heart	skeletal	NOTEST
ADAT3	chr19:1905372-1926012	heart	skeletal	NOTEST
ADCK4	chr19:41197433-41222790	heart	skeletal	NOTEST
AES	chr19:3052907-3062964	heart	skeletal	NOTEST
AKAP8	chr19:15464331-15490612	heart	skeletal	NOTEST
AKAP8L	chr19:15490858-15529833	heart	skeletal	NOTEST
AKT1	chr19:50377206-50380644	heart	skeletal	NOTEST

Filter (version 1.1.0)

Filter:

With following condition:

c14=='yes' and abs(c10)>1

Execute

test_id	gene_id	gene	locus
A1BG	A1BG	A1BG	chr19:58858171-58874214
A1BG-AS1	A1BG-AS1	A1BG-AS1	chr19:58858171-58874214
ABCA7	ABCA7	ABCA7	chr19:1040101-1065570

## 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/](http://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  $\log_{10}(\text{FPKM})$  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.

### ★Dendrogram

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

# Cuffdiff visualization with CummeRbund

## 7.1 Run CummeRbund tool

### ★CummeRbund

- Load the CummeRbund tool: NGS: RNA Analysis -> cummerbund
- 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: Dendrogram, check the “Replicates” box

- 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.

The screenshot shows the Galaxy web interface with the 'Galaxy / UMN' header. The left sidebar lists various NGS tools, including Picard, Assembly, Mapping, Indel Analysis, and RNA Analysis. The RNA Analysis section contains several sub-tools related to Cuffdiff and CummeRbund. A red 'a' highlights the 'cummeRbund' tool. The main workspace is titled 'cummeRbund (version 0.0.6)' and contains three plot configurations:

- Plots 1:** Set to 'Density' type, 'Replicates' checked, and 'Add new Plots' button circled in red.
- Plots 2:** Set to 'MultiDimensional Scaling (MDS) Plot' type, 'Replicates' checked, and 'Add new Plots' button circled in red.
- Plots 3:** Set to 'Dendrogram' type, 'Replicates' checked, and 'Add new Plots' button circled in red.

The right side of the screen shows a history of analysis steps, with a red 'b' pointing to the first step and a red 'c' pointing to the 'Execute' button at the bottom of the plot configuration pane.

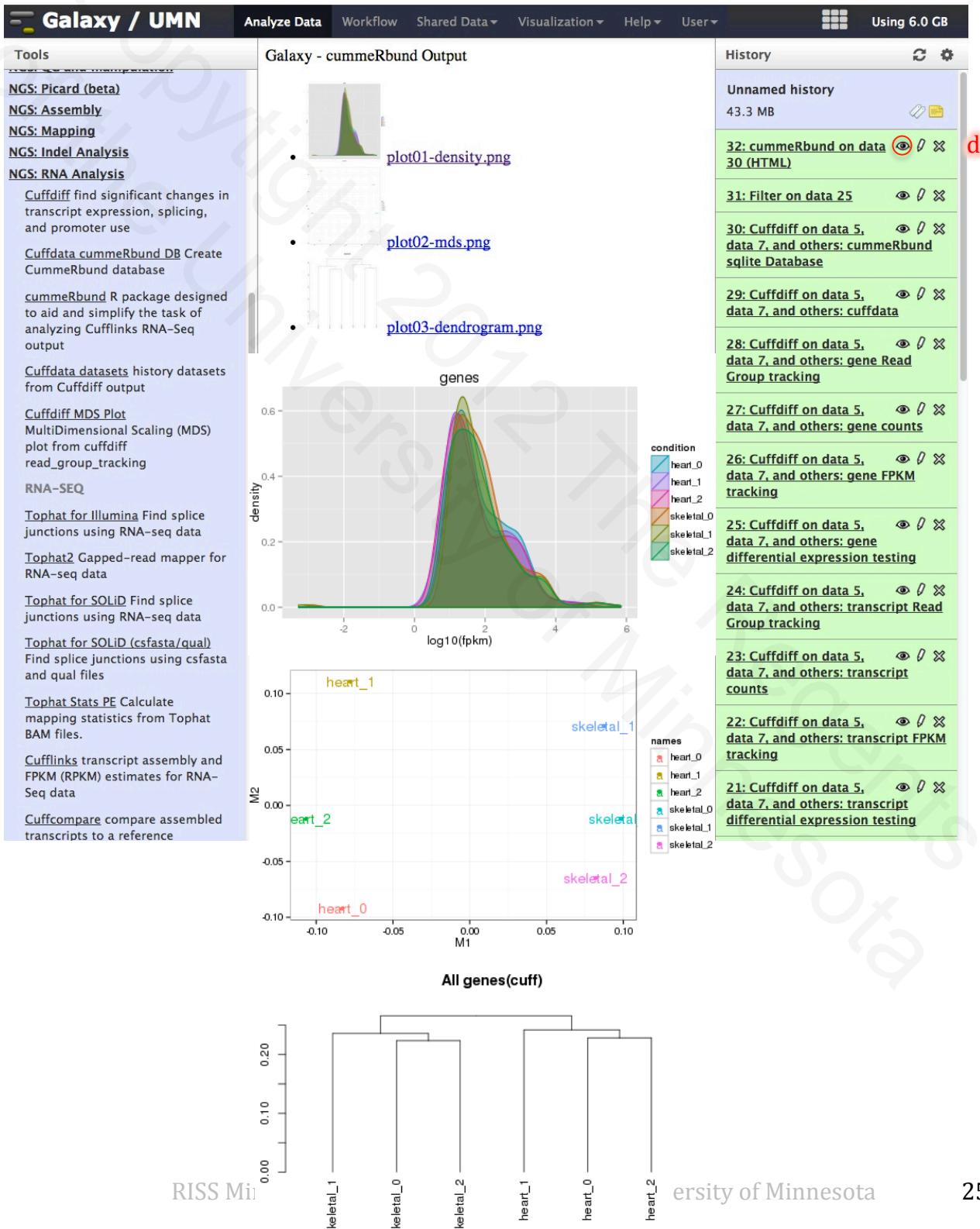
## Cuffdiff visualization with CummeRbund

### 7.2 Review CummeRbund plots

★Density plots, MDS plots, and Dendograms

- Click the “eye” icon next to the cummerbund output file to view the three plots
- Verify that:

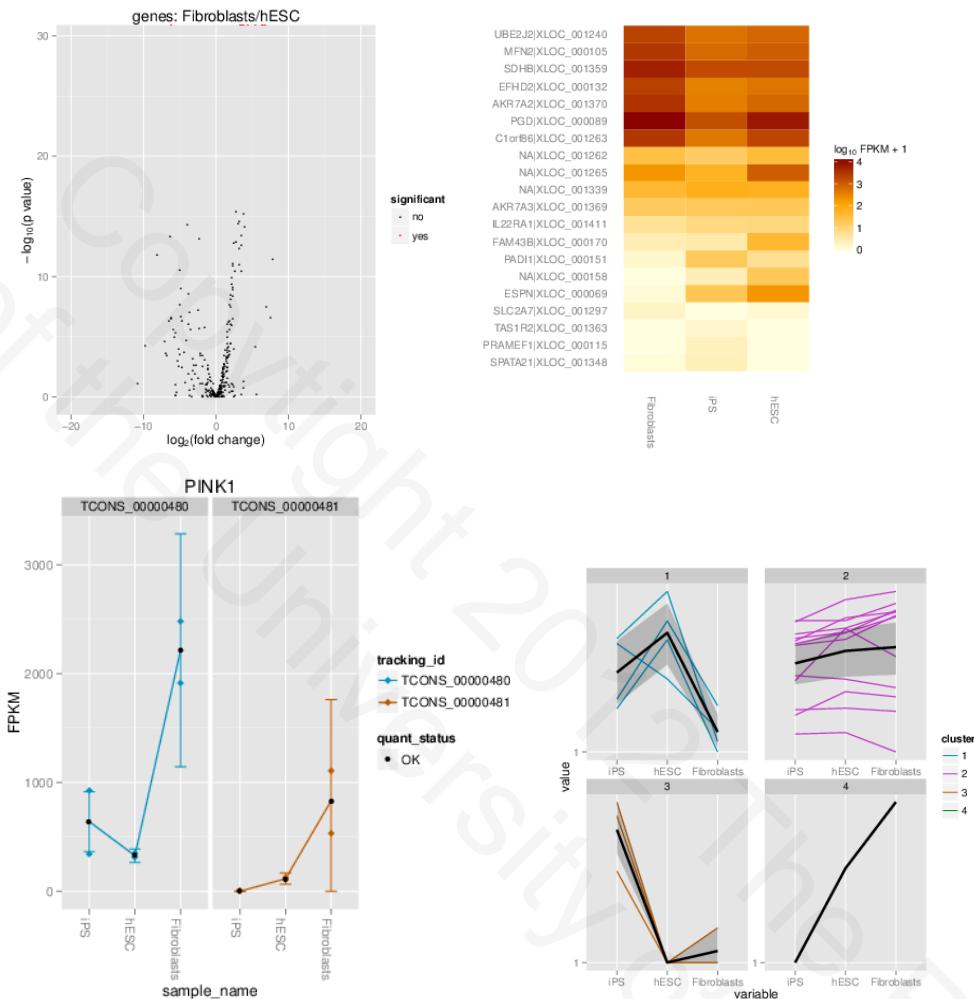
- The samples have similar density distributions
- The samples cluster by experimental condition in the MDS plot
- The sample cluster by experimental condition in the dendrogram



## Cuffdiff visualization with CummeRbund

### 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](mailto: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.

## Appendix A: Workflows

### 8.1 Extract workflow from current history

#### ★Galaxy Workflows

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

The following list contains each tool that was run to create the datasets in your current history. Please select those that you wish to include in the workflow.

Tools which cannot be run interactively and thus cannot be incorporated into a workflow will be shown in gray.

**Workflow name**

**Create Workflow**  Check all  Uncheck all

Tool	History items created
Unknown	1: heart-1_R1.fastq <input checked="" type="checkbox"/> Treat as input dataset
Unknown	2: heart-1_R2.fastq <input checked="" type="checkbox"/> Treat as input dataset
Unknown	3: hg19_chr19_genes_2012-03-09.gtf <input checked="" type="checkbox"/> Treat as input dataset
FastQC:Read QC	4: FastQC_heart-1_R1.fastq.html <input checked="" type="checkbox"/> Include "FastQC:Read QC" in workflow
Tophat2	5: Tophat2 on data 2, data 3, and data 1: align_summary 6: Tophat2 on data 2, data 3, and data 1: insertions 7: Tophat2 on data 2, data 3, and data 1: deletions 8: Tophat2 on data 2, data 3, and data 1: splice junctions 9: Tophat2 on data 2, data 3, and data 1: accepted_hits <input checked="" type="checkbox"/> Include "Tophat2" in workflow
Insertion size metrics	10: InsertSize_Insert size metrics.html <input checked="" type="checkbox"/> Include "Insert size metrics" in workflow
Tophat2	11: Tophat2 on data 2, data 3, and data 1: align_summary 12: Tophat2 on data 2, data 3, and data 1: insertions 13: Tophat2 on data 2, data 3, and data 1: deletions 14: Tophat2 on data 2, data 3, and data 1: splice junctions 15: Tophat2 on data 2, data 3, and data 1: accepted_hits <input type="checkbox"/> Include "Tophat2" in workflow

**History**

- 15: T data
- 14: T data
- 13: T data
- 12: T data
- 11: T data
- 10: I size
- 9: To 3, an
- 2.1 M forma
- Log: [2013-10-07 13:14:28] Checking for Samtools
- [2013-10-07 13:14:28] for Bowtie Bowtie version: 2.1.0.0 [2013-10-07 13:14:28] Checking for Samtools
- Import from File
- Dataset Security
- Resume Paused Jobs
- Collapse Expanded Datasets
- Include Deleted Datasets
- Include Hidden Datasets
- Unhide Hidden Datasets
- Purge Deleted Datasets
- Show Structure
- Export to File
- Delete
- Delete Permanently
- OTHER ACTIONS
- Import from File

Using 5.9 GB

## Appendix A: Workflows

### 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').

Galaxy / UMN

Analyze Data Workflow Shared Data Visualization Help User Using 5.9 GB

Your workflows

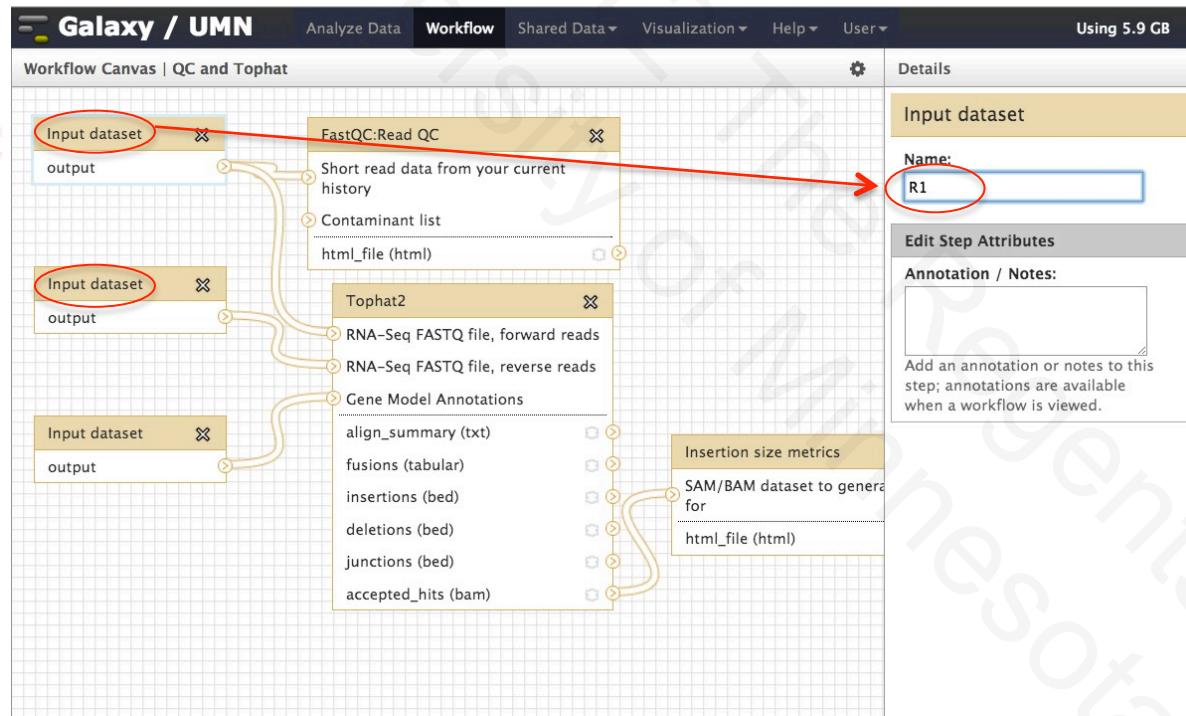
Name # of Steps

QC and Tophat	6
RN	18

Create new workflow Upload or import workflow

Workflow Context Menu:

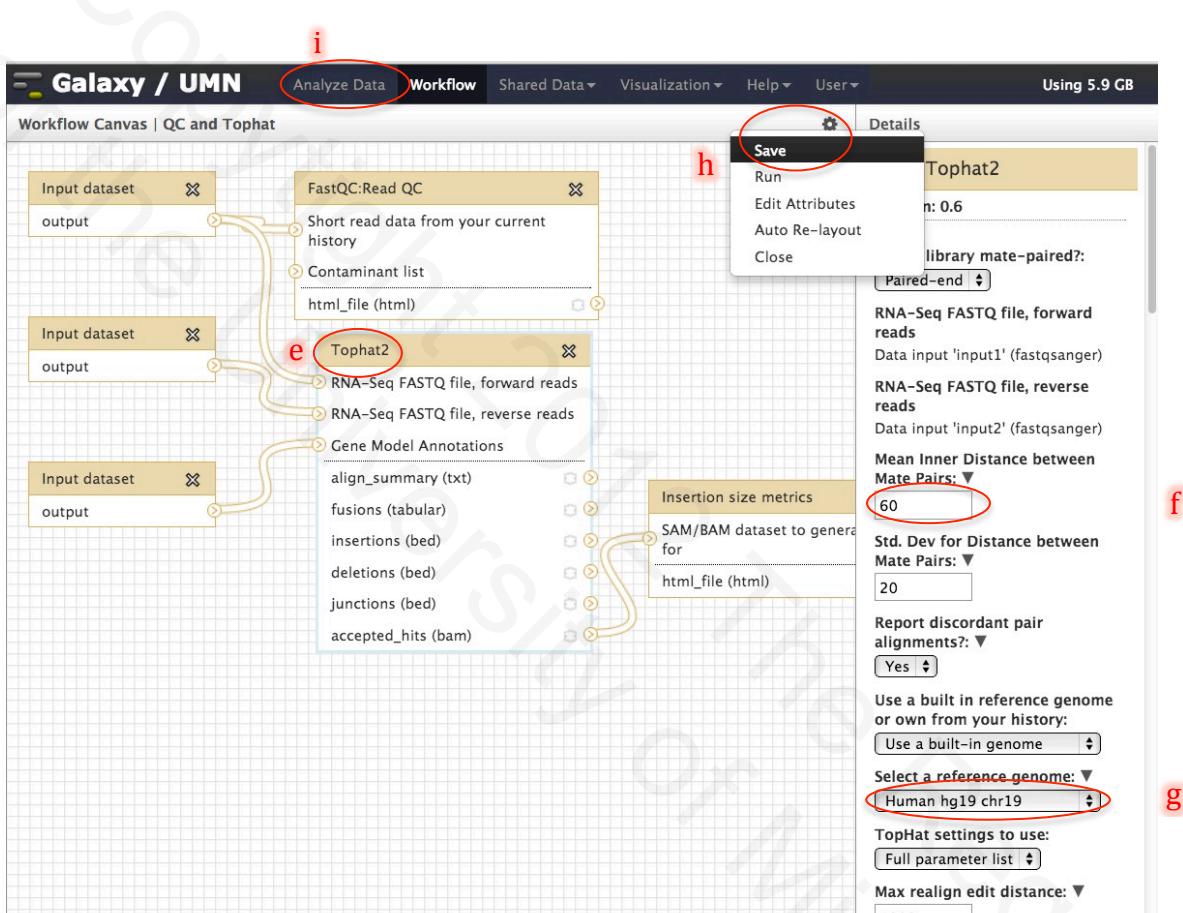
- Edit (highlighted)
- Run
- Share or Publish
- Download or Export
- Copy
- Rename
- View
- Delete



continued on next page...

## Appendix A: Workflows

- 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



## Appendix A: Workflows

### 8.3 Create new history

- 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.)
- 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
- Name the new history “heart-2”
- 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...*

The figure consists of three vertically stacked screenshots of the Galaxy web interface, showing the process of creating a new history and importing data from a data library.

- Screenshot 1:** Shows the main Galaxy interface with the Minnesota logo. The "History" panel on the right shows a history named "heart-1". A red circle labeled "a" highlights the "Create New" option in the dropdown menu that appears when the gear icon is clicked. Another red circle labeled "b" highlights the "Clone" option in the same menu.
- Screenshot 2:** Shows the main Galaxy interface. A red circle labeled "c" highlights the "heart-2" history in the "History" panel. A red circle labeled "d" highlights the "Data Libraries" link in the top navigation bar.
- Screenshot 3:** Shows the "Data Library 'RISS-tutorial-Hsapiens'" page. A red arrow points from the "RISS-tutorial-Hsapiens" link in Screenshot 2 to the "Fastq" section of this screenshot. Within the "Fastq" section, two files are selected: "heart-2\_R1.fastq" and "heart-2\_R2.fastq". A red circle labeled "e" highlights these selected files. A red arrow points from the "Import to current history" button at the bottom left of this screenshot to the "Go" button in Screenshot 1.

*continued on next page...*

## Appendix A: Workflows

- 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

**Screenshot (e): Data Library "RISS-tutorial-Hsapiens"**

The 'Shared Data' menu is open, showing the 'Data Libraries' option. The 'iGenomes' library is selected.

Name	Message	Data type	Date uploaded	File size
Bam	Bam alignment files			
Fastq	Raw RNA-seq fastq files			

**Screenshot (f): Data Library "iGenomes"**

The 'Analyze Data' menu is highlighted. The 'hg19\_chr19\_genes\_2012-03-09.gtf' file is selected. The 'Import to current history' button is highlighted.

Name	Message	Data type	Date uploaded	File size
GTF files	current GTF files for several organisms			
hg19_chr19_genes_2012-03-09.gtf	RNA-seq tutorial GTF file for Homo sapiens	gtf	2012-10-03	5.9 MB

For selected datasets:  Import to current history

**TIP:** You can download individual library datasets by selecting "Download this dataset" from the context menu (triangle) next to each dataset's name.  
**TIP:** Several compression options are available for downloading multiple library datasets simultaneously:  

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

## Appendix A: Workflows

### 8.4 Run workflow

- Load a workflow by clicking on “Workflow” at the top of the screen
- Click on the workflow that was just created and select “Run” from the dropdown menu
- 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
- Verify the GTF file is selected in the third drop-down menu
- Click on “Run workflow” to submit the FastQC, Tophat2, Insertion size metrics, and TophatstatsPE jobs.

The figure consists of three vertically stacked screenshots of the Galaxy / UMN web interface, each with red annotations:

- Screenshot 1 (Top):** Shows the main Galaxy header with "Workflow" highlighted (circled in red). The history panel shows two datasets.
- Screenshot 2 (Middle):** Shows the "Your workflows" page. A workflow named "QC and Tophat" is listed, with its context menu open. The "Run" option is highlighted and circled in red (annotation b).
- Screenshot 3 (Bottom):** Shows the "Running workflow 'QC and Tophat'" page. It displays four steps: Step 1: Input dataset (R1 dropdown), Step 2: Input dataset (Input Dataset dropdown), Step 3: Input dataset (Input Dataset dropdown), and Step 4: FastQC:Read QC (version 0.52). Annotations c, d, and e point to the dropdown menus for R1, the second input dataset, and the third input dataset respectively, all of which have their selected items circled in red. Annotation e also points to the "Run workflow" button at the bottom of the step 4 panel.