



WS9: RNA-Seq Analysis with Galaxy (non-model organism)

David Crossman, Ph.D.
UAB Heflin Center for Genomic Science

GCC2012
Wednesday, July 25, 2012

Galaxy Splash Page

The image shows a web browser window displaying the Galaxy web interface. The browser's address bar shows the URL `ec2-107-22-158-207.compute-1.amazonaws.com/#`. The browser's tab bar has two tabs, both labeled "Galaxy". The browser's bookmark bar shows several bookmarks, including "Galaxy", "Home | Integrative ...", "SEQanswers Home", "UAB My Drop Box", "Ingenuity - Ingenuit...", "Learn to code | Cod...", and "Welcome to ANGUS...". The browser's status bar shows "Using 29.7 Mb".

The Galaxy web interface has a dark blue header bar with the "Galaxy" logo on the left and navigation links "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User" on the right. The main content area is white and displays the text "Welcome to Galaxy on the Cloud" in a large, bold, black font. The left sidebar is light blue and contains a list of tool categories, each with a link to a page of tools. The right sidebar is light blue and contains a "History" section with a sub-header "Unnamed history" and "0 bytes". Below this, there is a message box that says "Your history is empty. Click 'Get Data' on the left pane to start".

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](#)
- [Graph/Display Data](#)
- [Regional Variation](#)
- [Multiple regression](#)
- [Multivariate Analysis](#)
- [Evolution](#)
- [Motif Tools](#)
- [Multiple Alignments](#)
- [Metagenomic analyses](#)
- [FASTA manipulation](#)
- [NCBI BLAST+](#)
- [NGS: QC and manipulation](#)
- [NGS: Picard](#)
- [NGS: Assembly](#)
- [NGS: Mapping](#)
- [NGS: Indel Analysis](#)
- [NGS: RNA Analysis](#)
- [NGS: SAM Tools](#)
- [NGS: GATK Tools](#)
- [NGS: Peak Calling](#)
- [SNP/WGA: Data; Filters](#)
- [SNP/WGA: QC; LD; Plots](#)
- [SNP/WGA: Statistical Models](#)
- [Human Genome Variation](#)
- [VCF Tools](#)

History

Unnamed history 0 bytes

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

Random Galaxy icons/colors

Colors



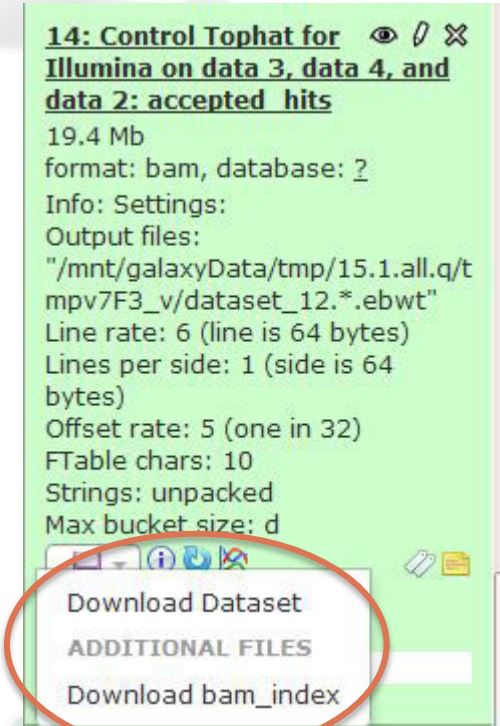
Queued

Running

Completed

Failed

Download/Save



Icons



Display data in browser



Edit attributes



Delete



Edit dataset annotation



View details



Run this job again



View in Trackster



Edit dataset tags

Edit files in History

Edit Attributes

Name:
Tophat for Illumina on data 3, data 4, and data 2 *

Info:
Settings:
Output files:

Annotation / Notes:
None
Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:
Click to Search or Select

Number of comment lines:

Chrom column:
1

Start column:
2

End column:
3




Strand column (click box & select):
☐ 1

Name/Identifier column (click box & select):
☐ 1

Score column for visualization:
1
2
3

Save

Auto-detect
This will inspect the dataset and attempt to correct the above column values if they are not accurate.

11: Tophat for Illumina   
on data 3, data 4, and data 2:
insertions

11: Control Tophat for   
Illumina on data 3, data 4, and
data 2: insertions

Upload/Import Data

Tools **1**

Get Data

- Upload File from your computer** **2**
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX main browser
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- Flymine test server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- metabolicMine server
- modENCODE worm server
- WormBase server
- Wormbase test server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- EpiGRAPH test server
- HbVar Human Hemoglobin Variants and Thalassemias

Upload File (version 1.1.3)

File Format:

Auto-detect **3a**

Which format? See help below

File:

Choose File No file chosen **3b-1**

TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

3b-2

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

File	Size	Date
<input type="checkbox"/> MF2_R1.fastqsanger	33.2 Mb	07/19/2012 07:26:42 AM
<input type="checkbox"/> MF2_R2.fastqsanger	33.2 Mb	07/19/2012 07:26:45 AM
<input type="checkbox"/> MF3_R1.fastqsanger	17.1 Mb	07/19/2012 07:26:47 AM
<input type="checkbox"/> MF3_R2.fastqsanger	17.1 Mb	07/19/2012 07:26:48 AM
<input type="checkbox"/> Treeshrew67 GeneScaffold_800_4487.gtf	17.3 Kb	07/19/2012 07:26:48 AM
<input type="checkbox"/> GeneScaffold_800_4487.fasta	251.2 Kb	07/19/2012 07:26:48 AM

3b-3

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at galaxy.uabgrid.uab.edu using your Galaxy credentials (email address and password).

Convert spaces to tabs:

☐ Yes

Use this option if you are entering intervals by hand.

Genome:

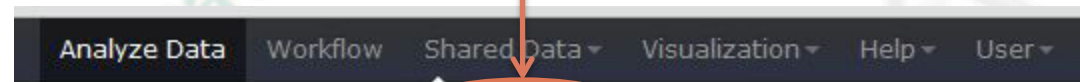
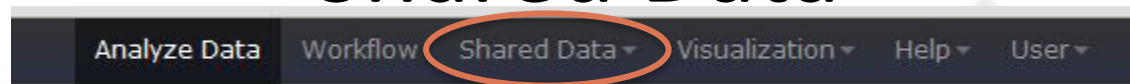
Click to Search or Select **3c**

Execute

3d

1. Click "Get Data"
2. Click "Upload File"
3. Boxes to be aware of:
 - a) File Format
 - b) File to be uploaded:
 - 1) File from computer
 - 2) URL/text
 - 3) FTP
 - c) Genome
4. Click "Execute"

Shared Data



3

Data Library "WS9: RNA-Seq Analysis with Galaxy"

Two exercises: mouse and treeshrew

Name	Message	Data type	Date uploaded	File size
mouse	FASTQ and .GTF files for a 2-condition (drugged X control) mouse experiment. Trimmed to just the genes Plekhh2-PigF [chr17:84911234..87424741]. Mouse data kindly provided by Kim Keeling, PhD at University of Alabama at Birmingham (http://www.microbio.uab.edu/faculty/keeling/index.html)			
	This dataset is uploading	None	2012-07-18	0 bytes
	Drugged sample, reverse reads	fastqsanger	2012-07-12	1.4 Mb
	Drugged sample, forward reads	fastqsanger	2012-07-12	1.4 Mb
	Control sample, forward reads	fastqsanger	2012-07-12	1.5 Mb
	Control sample, reverse reads	fastqsanger	2012-07-12	1.5 Mb
treeshrew	FASTQ, GTF, and FASTA files for a 2-condition (treated vs control) treeshrew experiment. Trimmed to only two scaffold regions (GeneScaffold_800 and GeneScaffold_4487).			
	GTF file from Ensembl for treeshrew version 67. Trimmed to only contain those genes in GeneScaffold_800 and GeneScaffold_4487.	gtf	2012-07-18	17.3 Kb
	FASTA file from Ensembl for treeshrew version 67. Trimmed to only contain those genes in GeneScaffold_800 and GeneScaffold_4487.	fasta	2012-07-18	251.2 Kb
	Control sample, reverse reads (R2)	fastqsanger	2012-07-18	33.2 Mb
	Control sample, forward reads (R1)	fastqsanger	2012-07-18	33.2 Mb
	Treated sample, reverse reads (R2)	fastqsanger	2012-07-18	17.1 Mb
	Treated sample, forward reads (R1)	fastqsanger	2012-07-18	17.1 Mb

1. Click on “Shared Data” (located on top toolbar)

2. Drop down box appears; click on “Data Libraries”

3. Will see this Data Library. Click on it to expand (as shown)

For selected datasets:

1. Click on "Shared Data" (located on top toolbar)
2. Drop down box appears; click on "Data Libraries"
3. Will see this Data Library. Click on it to expand (as shown)

Import Shared Data to Current History

1

<input checked="" type="checkbox"/>	treeshrew	FASTQ, GTF, and FASTA files for a 2-condition (treated vs control) treeshrew experiment. Trimmed to only two scaffold regions (GeneScaffold_800 and GeneScaffold_4487).			
<input checked="" type="checkbox"/>	Treeshrew67 GeneScaffold_800X_4487.gtf	GTF file from Ensembl for treeshrew version 67. Trimmed to only contain those genes in GeneScaffold_800 and GeneScaffold_4487.	gtf	2012-07-18	17.3 Kb
<input checked="" type="checkbox"/>	Treeshrew67_GeneScaffold_800_4487.fasta	FASTA file from Ensembl for treeshrew version 67. Trimmed to only contain those genes in GeneScaffold_800 and GeneScaffold_4487.	fasta	2012-07-18	251.2 Kb
<input checked="" type="checkbox"/>	Control R2.fastq	Control sample, reverse reads (R2)	fastqsanger	2012-07-18	33.2 Mb
<input checked="" type="checkbox"/>	Control R1.fastq	Control sample, forward reads (R1)	fastqsanger	2012-07-18	33.2 Mb
<input checked="" type="checkbox"/>	Treated R2.fastq	Treated sample, reverse reads (R2)	fastqsanger	2012-07-18	17.1 Mb
<input checked="" type="checkbox"/>	Treated R1.fastq	Treated sample, forward reads (R1)	fastqsanger	2012-07-18	17.1 Mb
For selected datasets: Import to current history <input type="button" value="Go"/> 2					

3

History	
Unnamed history	0 bytes
6: Treated R1.fastq	
5: Treated R2.fastq	
4: Control R1.fastq	
3: Control R2.fastq	
2: Treeshrew67 GeneScaffold_800_4487.fasta	
1: Treeshrew67 GeneScaffold_800X_4487.gtf	

1. Check boxes of files you want to import
2. Choose "Import to current history" and then click "Go"
3. Will see the files in the right-hand pane of the Galaxy window

Quality Control of raw fastq reads

Tools

- NGS: QC and manipulation** 1
- FASTQC: FASTQ/SAM/BAM
- Fastqc: Fastqc QC using FastQC from Babraham** 2
- ILLUMINA FASTQ
- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column
- ROCHE-454 DATA
- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

3a Fastqc: Fastqc QC (version 0.5)

Short read data from your current history:
3: Control R2.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 CAAGCAGAAGACGGCATACGA

Execute

3b Fastqc: Fastqc QC (version 0.5)

Short read data from your current history:
4: Control R1.fastq *

Title for the output file - to remind you what the job was for:
Control R1 FastQC *

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

Execute 4

1. Click on "NGS: QC and manipulation"
2. Click on "Fastqc: Fastqc QC"
3. Select options:
 - a) This is what the window looks like when first opened
 - b) Choose fastq file and give it a useful name
4. Click "Execute"
5. Do the exact same thing for the other 3 fastq files

FastQC Output Report

This data looks awful because this is filtered data from a much larger fastq file. Better results when using entire file!

History	
Unnamed history 2.7 Mb	
10: TreatedR2FastQC_data 5.html	
9: TreatedR1FastQC_data 6.html	
8: ControlR2FastQC_data 3.html	
7: ControlR1FastQC_data 4.html	
6: Treated R1.fastq	
5: Treated R2.fastq	
4: Control R1.fastq	
3: Control R2.fastq	
2: Treeshrew67_GeneScaffold_800_4487.fasta	
1: Treeshrew67_GeneScaffold_800X_4487.qtf	

Control_R1.fastq FastQC Report

FastQC Report

Thu 19 Jul 2012

Control_R1.fastq

Summary

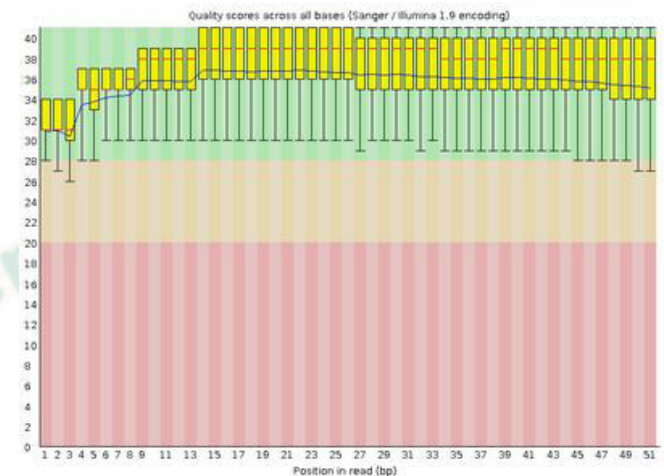
- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Basic Statistics

Measure	Value
Filename	Control_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	231776
Sequence length	51
%GC	46

[Back to summary](#)

Per base sequence quality



RNA-Seq Analysis Pipeline

TopHat

Aligns raw reads to reference genome, identifies splice junctions between exons.

Cufflinks

Assembles transcripts, and estimates their abundances.

Cuffmerge

Merges together several Cufflinks files.

Cuffdiff

Finds significant changes in transcript expression, splicing, and promoter use.

TopHat

1 NGS: RNA Analysis

2 RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use
- Cuffmerge merge together several Cufflinks assemblies

FILTERING

- Filter Combined Transcripts using tracking file

3 Tophat for Illumina (version 1.5.0)

RNA-Seq FASTQ file:
6: Treated R1.fastq

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Will you select a reference genome from your history or use a built-in index?:
Use a built-in index

Built-ins were indexed using default options

Select a reference genome:
(Zea mays): Zea_mays_B73_RefGe

If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:
Single-end

TopHat settings to use:
Default settings

Use the Full parameter list to change default settings.

Execute

1. Click on “NGS: RNA Analysis”
2. Click on “Tophat for Illumina”
3. Default window with options appears

TopHat

Tophat for Illumina (version 1.5.0)

RNA-Seq FASTQ file:

4: Control R1.fastq ▼ 1

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Will you select a reference genome from your history or use a built-in index?:

Use one from the history ▼ 2a

Built-ins were indexed using default options

Select the reference genome:

2: Treeshrew67_GeneS.._4487.fasta ▼ 2b

Is this library mate-paired?:

Paired-end ▼ 3

RNA-Seq FASTQ file:

3: Control R2.fastq ▼ 4

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs:

150 5

TopHat settings to use:

Default settings ▼ 6

Use the Full parameter list to change default settings.

Execute 7

1. Select forward fastq read file
2. Select reference genome:
 - a) Choose "Use one from the history"
 - b) Select the reference genome fasta
3. Select "Paired-end"
4. Select reverse fastq read file
5. Input "150" (ask sequencing center for this info)
6. Can choose "Default settings" or "Full parameter list"
7. Click "Execute"
8. Do the exact same thing for the other sample

Note about FASTA files not already indexed in Galaxy

- If a FASTA is not indexed in Galaxy, then it is easy to upload the appropriate FASTA file into Galaxy. (Get Data -> Upload File)
- However, it can take up to 5 hours extra to run TopHat because Bowtie has to index your uploaded FASTA file (best to have your own instance of Galaxy) each time you run TopHat!
- Where do I go to get a non-model organism FASTA file?
 - NCBI: <http://www.ncbi.nlm.nih.gov/genome>
 - Ensembl: <http://useast.ensembl.org/info/data/ftp/index.html>
 - iGenome: <http://cufflinks.cbcb.umd.edu/igenomes.html>
 - Your favorite species website: <http://www...>

TopHat output files



The following job has been successfully added to the queue:

11: Tophat for Illumina on data 3, data 4, and data 2: insertions

12: Tophat for Illumina on data 3, data 4, and data 2: deletions

13: Tophat for Illumina on data 3, data 4, and data 2: splice junctions

14: Tophat for Illumina on data 3, data 4, and data 2: accepted_hits

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History		
Unnamed history		32.3 Mb
18: Treated Tophat for Illumina on data 5, data 6, and data 2: accepted_hits	👁	🗑
17: Treated Tophat for Illumina on data 5, data 6, and data 2: splice junctions	👁	🗑
16: Treated Tophat for Illumina on data 5, data 6, and data 2: deletions	👁	🗑
15: Treated Tophat for Illumina on data 5, data 6, and data 2: insertions	👁	🗑
14: Control Tophat for Illumina on data 3, data 4, and data 2: accepted_hits	👁	🗑
13: Control Tophat for Illumina on data 3, data 4, and data 2: splice junctions	👁	🗑
12: Control Tophat for Illumina on data 3, data 4, and data 2: deletions	👁	🗑
11: Control Tophat for Illumina on data 3, data 4, and data 2: insertions	👁	🗑

Cufflinks

NGS: RNA Analysis

RNA-SEQ

- [Tophat for Illumina](#) Find splice junctions using RNA-seq data
- [Cufflinks](#) transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- [Cuffcompare](#) compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- [Cuffdiff](#) find significant changes in transcript expression, splicing, and promoter use
- [Cuffmerge](#) merge together several Cufflinks assemblies

FILTERING

- [Filter Combined Transcripts](#) using tracking file

3 Cufflinks (version 0.0.5)

SAM or BAM file of aligned RNA-Seq reads:

18: Treated Tophat fo...epted_hits ▾

Max Intron Length:

300000

Min Isoform Fraction:

0.1

Pre mRNA Fraction:

0.15

Perform quartile normalization:

No ▾

Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Use Reference Annotation:

No ▾

Perform Bias Correction:

No ▾

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Set Parameters for Paired-end Reads? (not recommended):

No ▾

Execute

1. Click on “NGS: RNA Analysis”
2. Click on “Cufflinks”
3. Default window with options appears

Cufflinks

Cufflinks (version 0.0.5)

SAM or BAM file of aligned RNA-Seq reads:

14: Control Tophat fo...cepted_hits **1**

Max Intron Length:

300000

Min Isoform Fraction:

0.1

Pre MRNA Fraction:

0.15

Perform quartile normalization:

No **2**

Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Use Reference Annotation:

Use reference annotation **3a**

Reference Annotation:

1: Treeshrew67 GeneS..0X_4487.gtf **3b**

Gene annotation dataset in GTF or GFF3 format.

Perform Bias Correction:

No **4**

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Set Parameters for Paired-end Reads? (not recommended):

No

1. Choose TopHat accepted hits file
2. Perform quartile normalization (for this demo sample, choose "No")
3. Reference Annotation:
 - a) For genomes in scaffolds, choose "Use reference annotation"
 - b) Choose GTF file from history
4. Perform Bias Correction (for this demo, choose "No")
5. Click "Execute"
6. Do the exact same thing for the other TopHat accepted hits file

Note about GTF files for Cuff*

- If you use a GTF file from Ensembl, then you need to convert the chromosome column (column 1) to include 'chr' in front of the chromosome #. You can do this by:
 - Using Jeremy's published workflow "Make Ensembl GTF compatible with Cufflinks" in Galaxy:
<https://main.g2.bx.psu.edu/u/jeremy/w/make-ensembl-gtf-compatible-with-cufflinks>
 - Use 'awk' to add 'chr' to column 1 (if using Mac or Linux)
- Where do I go to get a GTF file?
 - NCBI: <http://www.ncbi.nlm.nih.gov/genome>
 - Ensembl: <http://useast.ensembl.org/info/data/ftp/index.html>
 - iGenome: <http://cufflinks.cbcb.umd.edu/igenomes.html>
 - Your favorite species website: <http://www...>

Cufflinks output files



The following job has been successfully added to the queue:

23: Cufflinks on data 18 and data 1: gene expression

24: Cufflinks on data 18 and data 1: transcript expression

25: Cufflinks on data 18 and data 1: assembled transcripts

26: Cufflinks on data 18 and data 1: total map mass

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History	
Unamed history 32.3 Mb	
<u>25: Treated Cufflinks on data 18 and data 1: assembled transcripts</u>	👁️ 🗑️ ✕
<u>24: Treated Cufflinks on data 18 and data 1: transcript expression</u>	👁️ 🗑️ ✕
<u>23: Treated Cufflinks on data 18 and data 1: gene expression</u>	👁️ 🗑️ ✕
<u>21: Control Cufflinks on data 14 and data 1: assembled transcripts</u>	👁️ 🗑️ ✕
<u>20: Control Cufflinks on data 14 and data 1: transcript expression</u>	👁️ 🗑️ ✕
<u>19: Control Cufflinks on data 14 and data 1: gene expression</u>	👁️ 🗑️ ✕

Cuffmerge

NGS: RNA Analysis

RNA-SEQ

- [Tophat for Illumina](#) Find splice junctions using RNA-seq data
- [Cufflinks](#) transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- [Cuffcompare](#) compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- [Cuffdiff](#) find significant changes in transcript expression, splicing, and promoter use
- [Cuffmerge](#) merge together several Cufflinks assemblies

FILTERING

- [Filter Combined Transcripts](#) using tracking file

Cuffmerge (version 0.0.5)

GTF file produced by Cufflinks:

25: Treated Cufflinks..transcripts

Additional GTF Input Files

Add new Additional GTF Input Files

Use Reference Annotation:

No

Use Sequence Data:

No

Use sequence data for some optional classification functions, including the addition of the p_id attribute required by Cuffdiff.

Execute

1. Click on “NGS: RNA Analysis”
2. Click on “Cuffmerge”
3. Default window with options appears

Cuffmerge

Cuffmerge (version 0.0.5)

GTF file produced by Cufflinks:

21: Control Cufflinks..transcripts

1

Additional GTF Input Files

Additional GTF Input Files 1

GTF file produced by Cufflinks:

25: Treated Cufflinks..transcripts

2b

Remove Additional GTF Input Files 1

Add new Additional GTF Input Files

2a

Use Reference Annotation:

Yes

3a

Reference Annotation:

1: Treeshrew67 GeneS..0X_4487.gtf

3b

Make sure your annotation file is in GTF format and that Galaxy knows that your file is GTF--not GFF.

Use Sequence Data:

Yes

4a

Use sequence data for some optional classification functions, including the addition of the p_id attribute required by Cuffdiff.

Choose the source for the reference list:

History

4b

Using reference file:

2: Treeshrew67_GeneS.._4487.fasta

4c

Execute

5

1. Choose GTF file produced by Cufflinks
2. Additional GTF Input Files:
 - a) Click on "Add new Additional GTF Input Files"
 - b) Choose other GTF file produced by Cufflinks
3. Reference Annotation:
 - a) Select "Yes" to Use Reference Annotation
 - b) Choose GTF Reference Annotation file from history
4. Sequence Data:
 - a) Select "Yes" to Use Sequence Data
 - b) Choose "History"
 - c) Choose FASTA file from history
5. Click "Execute"

Cuffmerge output files



The following job has been successfully added to the queue:

27: Cuffmerge on data 21, data 1, and others: merged transcripts

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History

Unnamed history 32.3 Mb

27: Treated vs Control

Cuffmerge on data 21, data 1, and others: merged transcripts

Cuffdiff

NGS: RNA Analysis

RNA-SEQ

- [Tophat for Illumina](#) Find splice junctions using RNA-seq data
- [Cufflinks](#) transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- [Cuffcompare](#) compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- [Cuffdiff](#) find significant changes in transcript expression, splicing, and promoter use
- [Cuffmerge](#) merge together several Cufflinks assemblies

FILTERING

- [Filter Combined Transcripts](#) using tracking file

Cuffdiff (version 0.0.5)

Transcripts:

27: Treated vs Contro..transcripts

A transcript GTF file produced by cufflinks, cuffcompare, or other source.

Perform replicate analysis:

No

Perform cuffdiff with replicates in each group.

SAM or BAM file of aligned RNA-Seq reads:

18: Treated Tophat fo..cepted_hits

SAM or BAM file of aligned RNA-Seq reads:

18: Treated Tophat fo..cepted_hits

False Discovery Rate:

0.05

The allowed false discovery rate.

Min Alignment Count:

10

The minimum number of alignments in a locus for needed to conduct significance testing on changes in that locus observed between samples.

Perform quartile normalization:

No

Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Perform Bias Correction:

No

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Set Parameters for Paired-end Reads? (not recommended):

No

Execute

1. Click on "NGS: RNA Analysis"
2. Click on "Cuffdiff"
3. Default window with options appears

Cuffdiff

Cuffdiff (version 0.0.5)

Transcripts:
 27: Treated vs Contro..transcripts 1
 A transcript GTF file produced by cufflinks, cuffcompare, or other source.

Perform replicate analysis:
 Yes 2a
 Perform cuffdiff with replicates in each group.

Groups

Group 1
 Group name (no spaces or commas):
 Control 2c

Replicates
 Replicate 1
 Add file:
 14: Control Tophat fo...cepted_hits 2d
 Remove Replicate 1

Add new Replicate 2e
 Remove Group 1

Group 2
 Group name (no spaces or commas):
 Treated 2g

Replicates
 Replicate 1
 Add file:
 18: Treated Tophat fo...cepted_hits 2h
 Remove Replicate 1

Add new Replicate 2i
 Remove Group 2

Add new Group 2b, 2f, 2j

False Discovery Rate:
 0.05 3
 The allowed false discovery rate.

Min Alignment Count:
 10 4
 The minimum number of alignments in a locus for needed to conduct significance testing on changes in that locus observed between samples.

Perform quartile normalization:
 No 5
 Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Perform Bias Correction:
 No 6
 Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Set Parameters for Paired-end Reads? (not recommended):
 No

Execute 7

1. Choose GTF transcript file from either Cuffmerge or Cuffcompare
2. Perform replicate analysis:
 - a) Choose "Yes"
 - b) Click "Add new Group"
 - c) Select a name to give the Group
 - d) Choose TopHat accepted hits file associated with this Group
 - e) If you have more than one TopHat accepted hits file associated with this Group, then click "Add new Replicate"
 - f) Click "Add new Group"
 - g) Select a name to give the Group
 - h) Choose TopHat accepted hits file associated with this Group
 - i) If you have more than one TopHat accepted hits file associated with this Group, then click "Add new Replicate"
 - j) Click "Add new Group" if you have another Group you want to add
3. Select a False Discovery Rate cutoff
4. Select the minimum # of reads that will align to a locus in order to perform significant testing
5. Perform quartile normalization (for this demo, choose "No")
6. Perform bias correction (for this demo, choose "No")
7. Click "Execute"

Cuffdiff output files



The following job has been successfully added to the queue:

- 28: Cuffdiff on data 18, data 14, and data 27: splicing differential expression testing
- 29: Cuffdiff on data 18, data 14, and data 27: promoters differential expression testing
- 30: Cuffdiff on data 18, data 14, and data 27: CDS overloading differential expression testing
- 31: Cuffdiff on data 18, data 14, and data 27: CDS FPKM differential expression testing
- 32: Cuffdiff on data 18, data 14, and data 27: CDS FPKM tracking
- 33: Cuffdiff on data 18, data 14, and data 27: TSS groups differential expression testing
- 34: Cuffdiff on data 18, data 14, and data 27: TSS groups FPKM tracking
- 35: Cuffdiff on data 18, data 14, and data 27: gene differential expression testing
- 36: Cuffdiff on data 18, data 14, and data 27: gene FPKM tracking
- 37: Cuffdiff on data 18, data 14, and data 27: transcript differential expression testing
- 38: Cuffdiff on data 18, data 14, and data 27: transcript FPKM tracking

You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

History	
Unnamed history 32.3 Mb	
38: Cuffdiff on data 18, data 14, and data 27: transcript FPKM tracking	🔍 🗑️
37: Cuffdiff on data 18, data 14, and data 27: transcript differential expression testing	🔍 🗑️
36: Cuffdiff on data 18, data 14, and data 27: gene FPKM tracking	🔍 🗑️
35: Cuffdiff on data 18, data 14, and data 27: gene differential expression testing	🔍 🗑️
34: Cuffdiff on data 18, data 14, and data 27: TSS groups FPKM tracking	🔍 🗑️
33: Cuffdiff on data 18, data 14, and data 27: TSS groups differential expression testing	🔍 🗑️
32: Cuffdiff on data 18, data 14, and data 27: CDS FPKM tracking	🔍 🗑️
31: Cuffdiff on data 18, data 14, and data 27: CDS FPKM differential expression testing	🔍 🗑️
30: Cuffdiff on data 18, data 14, and data 27: CDS overloading differential expression testing	🔍 🗑️
29: Cuffdiff on data 18, data 14, and data 27: promoters differential expression testing	🔍 🗑️
28: Cuffdiff on data 18, data 14, and data 27: splicing differential expression testing	🔍 🗑️

Transcript differential expression testing output

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value
TCONS_00000001	XLOC_000001	MRPL52	GeneScaffold_4487:152198-156919	Control	Treated	OK	136514	109554	-0.317401	2.94565	0.00322275	0.00537126
TCONS_00000002	XLOC_000002	MMP14	GeneScaffold_4487:164659-168821	Control	Treated	OK	127591	419087	1.71572	-51.7058	0	0
TCONS_00000003	XLOC_000003	SLC7A7	GeneScaffold_4487:2-126817	Control	Treated	OK	2810.33	1640.05	-0.776996	2.29497	0.0217349	0.0271686
TCONS_00000004	XLOC_000004	ENSTBEG00000000939	GeneScaffold_4487:2-126817	Control	Treated	NOTEST	0	0	0	0	1	1
TCONS_00000005	XLOC_000005	ENSTBEG00000000949	GeneScaffold_4487:2-126817	Control	Treated	OK	799.391	730.767	-0.12949	0.0897554	0.928482	0.928482
TCONS_00000006	XLOC_000006	TIMP3	GeneScaffold_800:2226-57391	Control	Treated	OK	1.63872e+06	764572	-1.09984	44.4146	0	0

Gene differential expression testing output

test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	sig
XLOC_000001	XLOC_000001	MRPL52	GeneScaffold_4487:152198-156919	Control	Treated	OK	136514	109554	-0.317401	2.94565	0.00322275	0.00537126	ye
XLOC_000002	XLOC_000002	MMP14	GeneScaffold_4487:164659-168821	Control	Treated	OK	127591	419087	1.71572	-51.7058	0	0	ye
XLOC_000003	XLOC_000003	SLC7A7	GeneScaffold_4487:2-126817	Control	Treated	OK	2810.33	1640.05	-0.776996	2.29497	0.0217349	0.0271686	ye
XLOC_000004	XLOC_000004	ENSTBEG00000000939	GeneScaffold_4487:2-126817	Control	Treated	NOTEST	0	0	0	0	1	1	no
XLOC_000005	XLOC_000005	ENSTBEG00000000949	GeneScaffold_4487:2-126817	Control	Treated	OK	799.391	730.767	-0.12949	0.0897554	0.928482	0.928482	no
XLOC_000006	XLOC_000006	TIMP3	GeneScaffold_800:2226-57391	Control	Treated	OK	1.63872e+06	764572	-1.09984	44.4146	0	0	ye

Create genome in IGV

The screenshot shows the IGV application window. The 'File' menu is open, and the 'Import Genome...' option is highlighted. The interface includes a menu bar (File, View, Tracks, Regions, Help), a search bar with a 'Go' button, a chromosome scale (2-19, X), a main track area, and a 'Refseq genes' track at the bottom. A status bar at the bottom right indicates '82M of 259M'.

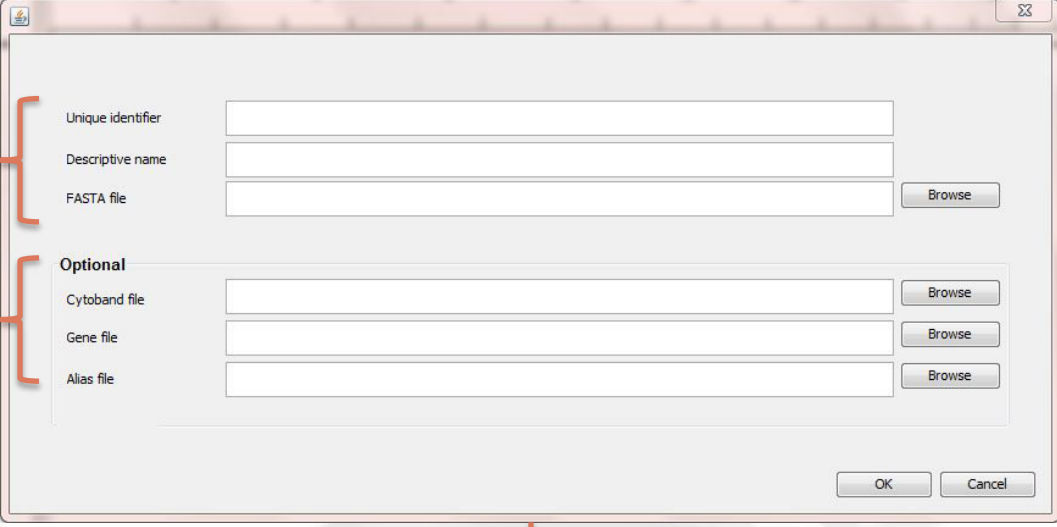
1. Click on "File"

2. Choose "Import Genome"

Create genome in IGV

Required

Optional



The dialog box is empty, showing the following fields and buttons:

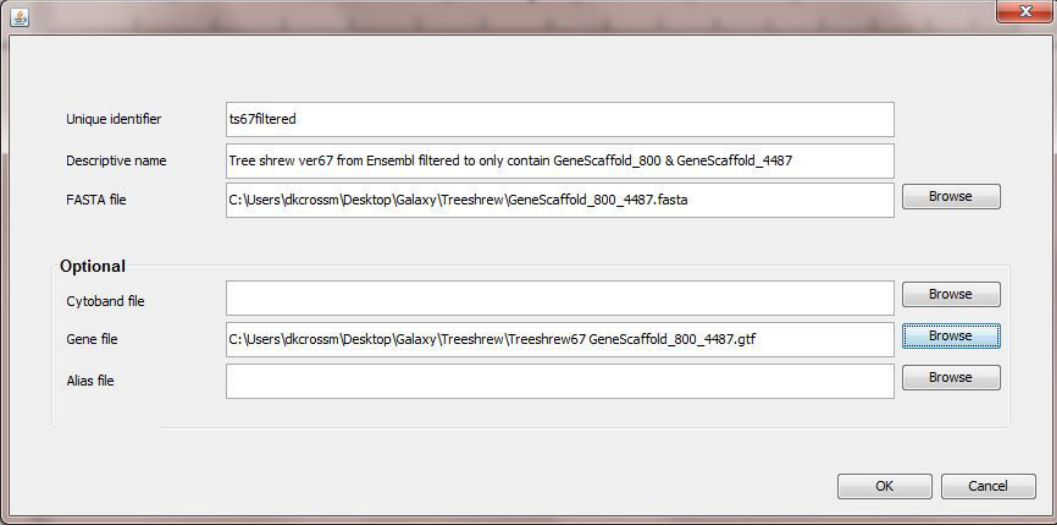
- Unique identifier:
- Descriptive name:
- FASTA file:
- Optional**
 - Cytoband file:
 - Gene file:
 - Alias file:
-

Need:

1. FASTA file (required)

Optional:

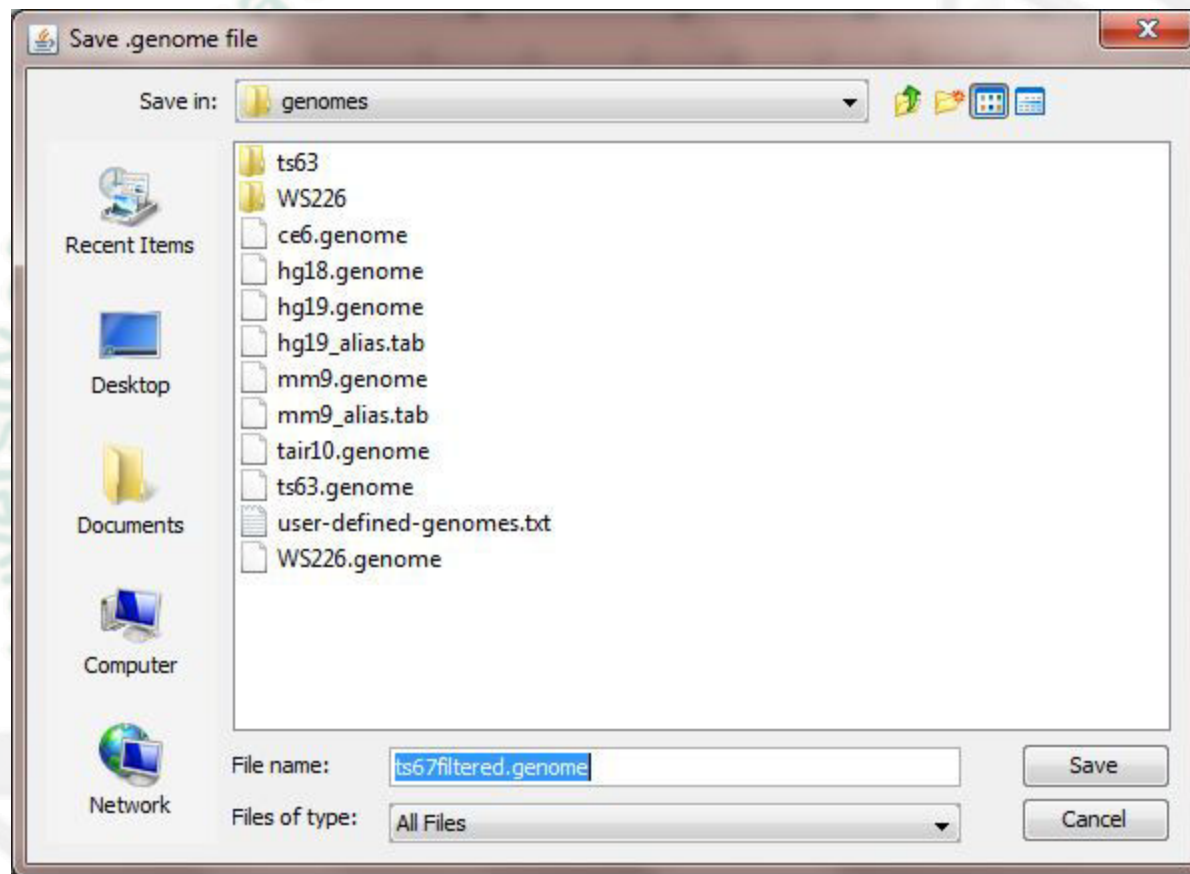
1. Cytoband file
2. Gene file (can use GTF)
3. Alias file



The dialog box is filled with the following data:

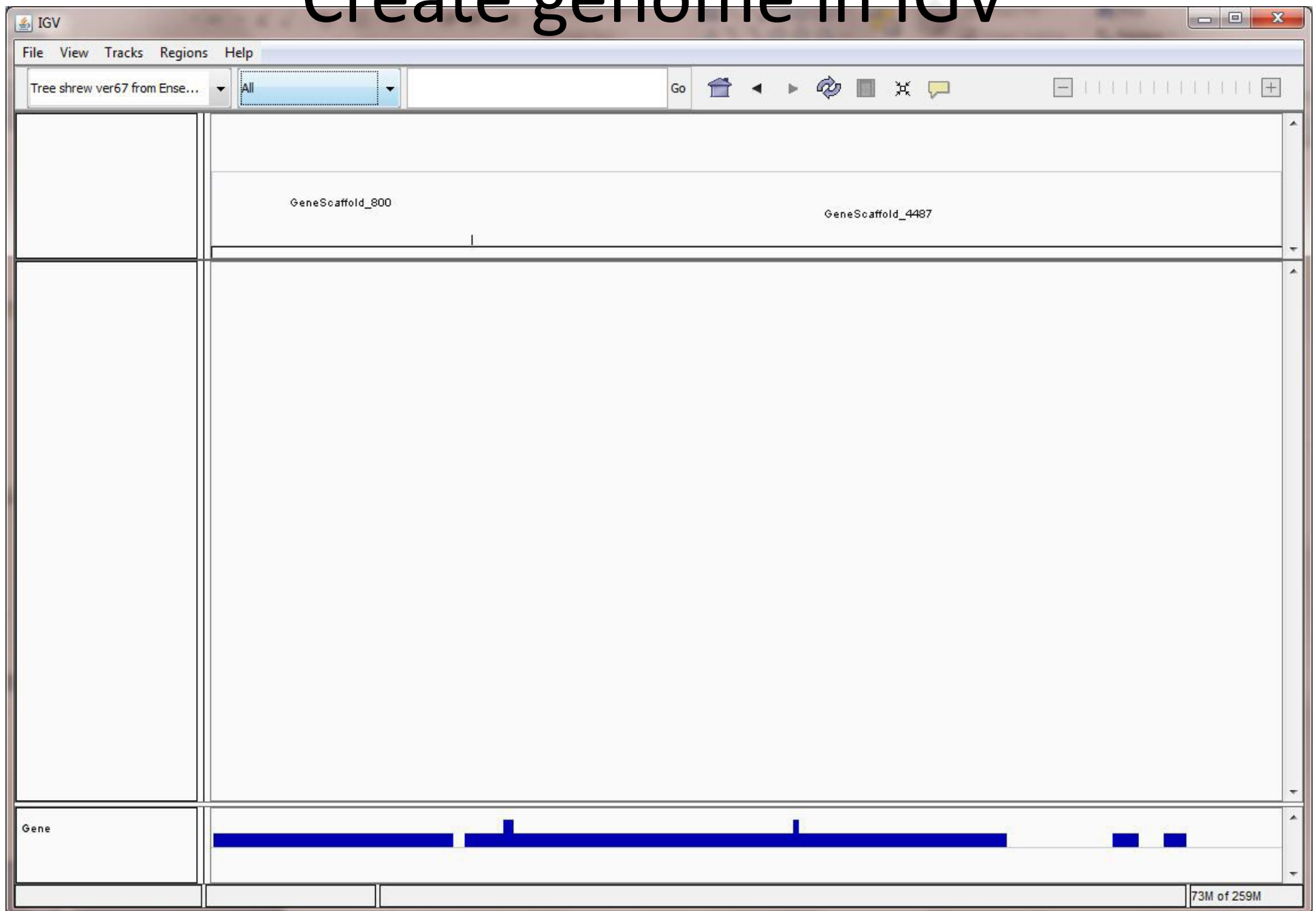
- Unique identifier: ts67filtered
- Descriptive name: Tree shrew ver67 from Ensembl filtered to only contain GeneScaffold_800 & GeneScaffold_4487
- FASTA file: C:\Users\dkcrossm\Desktop\Galaxy\Treeshrew\GeneScaffold_800_4487.fasta
- Optional**
 - Cytoband file:
 - Gene file: C:\Users\dkcrossm\Desktop\Galaxy\Treeshrew\Treeshrew67 GeneScaffold_800_4487.gtf
 - Alias file:
-

Create genome in IGV



Save the
*.genome
file

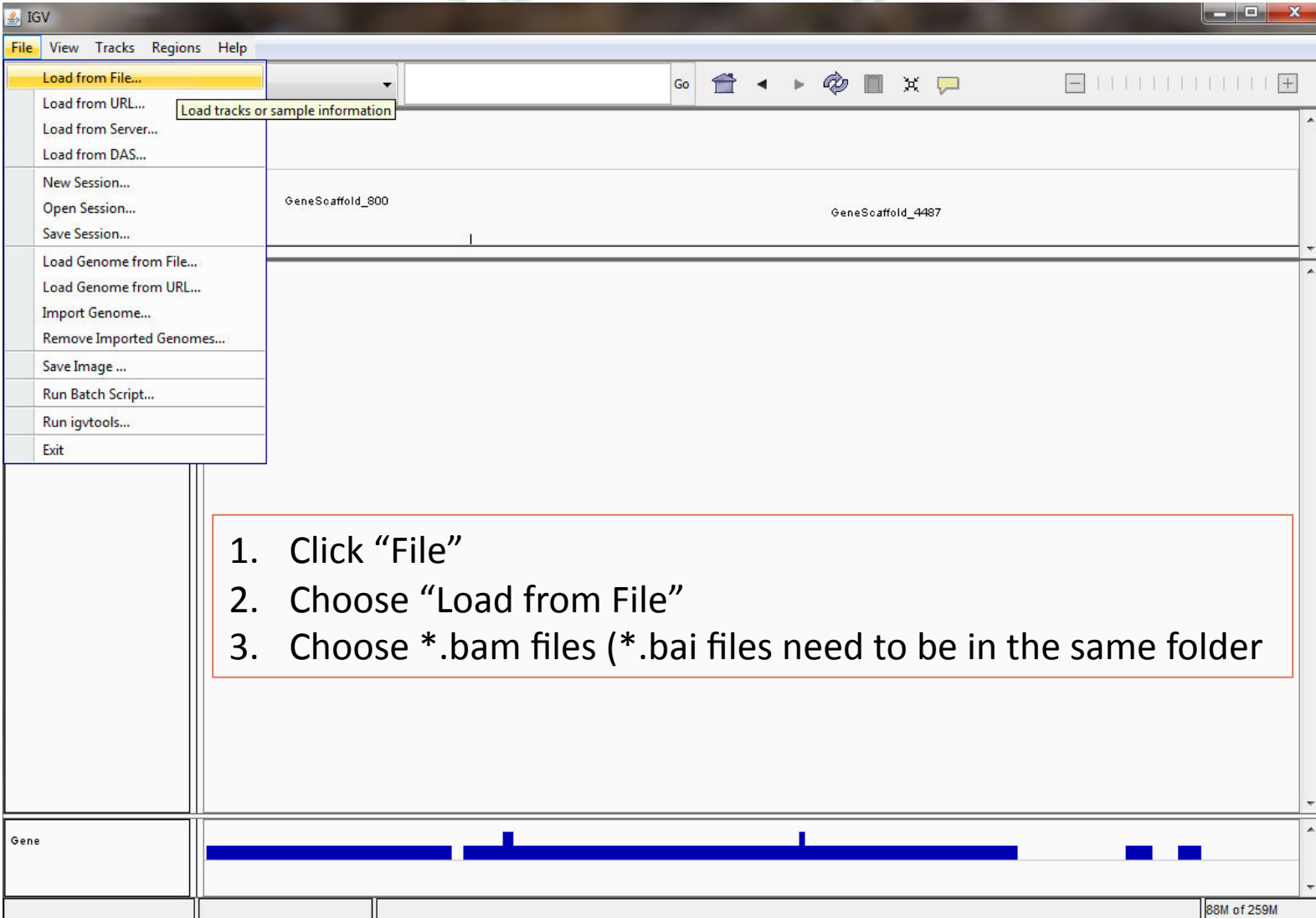
Create genome in IGV



Load aligned BAM files into IGV

1

2



IGV

File View Tracks Regions Help

Load from File... Load tracks or sample information

Load from URL...

Load from Server...

Load from DAS...

New Session...

Open Session...

Save Session...

Load Genome from File...

Load Genome from URL...

Import Genome...

Remove Imported Genomes...

Save Image ...

Run Batch Script...

Run igvtools...

Exit

GeneScaffold_800

GeneScaffold_4487

1. Click "File"

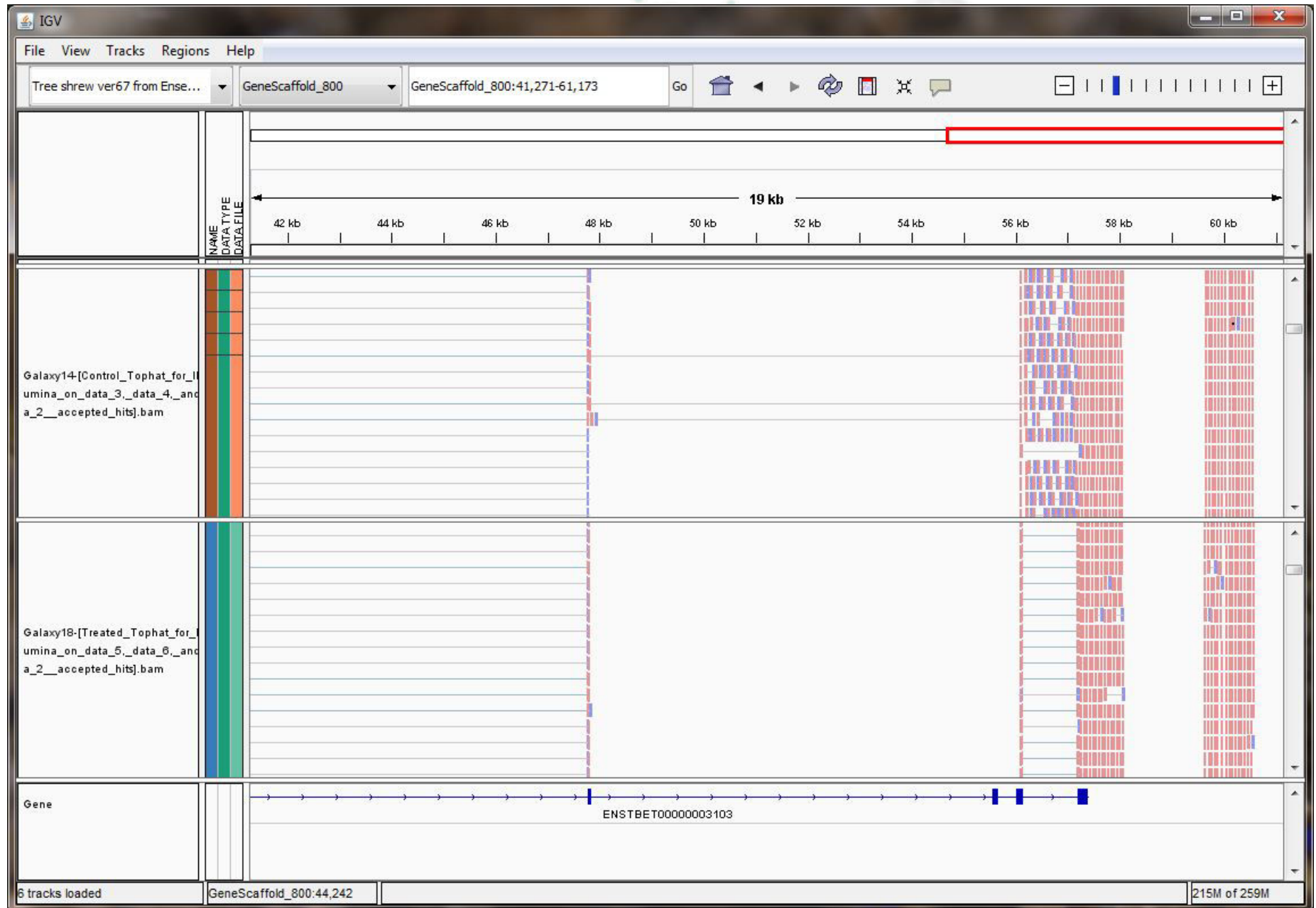
2. Choose "Load from File"

3. Choose *.bam files (*.bai files need to be in the same folder)

Gene

88M of 259M

IGV



References and web links

- TopHat
 - Trapnell C, Pachter L, Salzberg SL. [TopHat: discovering splice junctions with RNA-Seq](#). *Bioinformatics* doi:10.1093/bioinformatics/btp120
 - <http://tophat.cbcb.umd.edu/>
- Bowtie
 - Langmead B, Trapnell C, Pop M, Salzberg SL. [Ultrafast and memory-efficient alignment of short DNA sequences to the human genome](#). *Genome Biol* 10:R25.
 - <http://bowtie-bio.sourceforge.net/index.shtml>
- Cufflinks
 - Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. [Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation](#) *Nature Biotechnology* doi:10.1038/nbt.1621
 - Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. [Improving RNA-Seq expression estimates by correcting for fragment bias](#) *Genome Biology* doi:10.1186/gb-2011-12-3-r22
 - Roberts A, Pimentel H, Trapnell C, Pachter L. [Identification of novel transcripts in annotated genomes using RNA-Seq](#) *Bioinformatics* doi:10.1093/bioinformatics/btr355
 - <http://cufflinks.cbcb.umd.edu/>
- TopHat and Cufflinks protocol
 - Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. [Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks](#) *Nature Protocols* 7, 562-578 (2012) doi:10.1038/nprot.2012.016
- IGV
 - <http://www.broadinstitute.org/igv/>

Thanks! Questions?

Contact info:

David K. Crossman, Ph.D.

Bioinformatics Director

Heflin Center for Genomic Science

University of Alabama at Birmingham

<http://www.heflingenetics.uab.edu>

dkcrossm@uab.edu