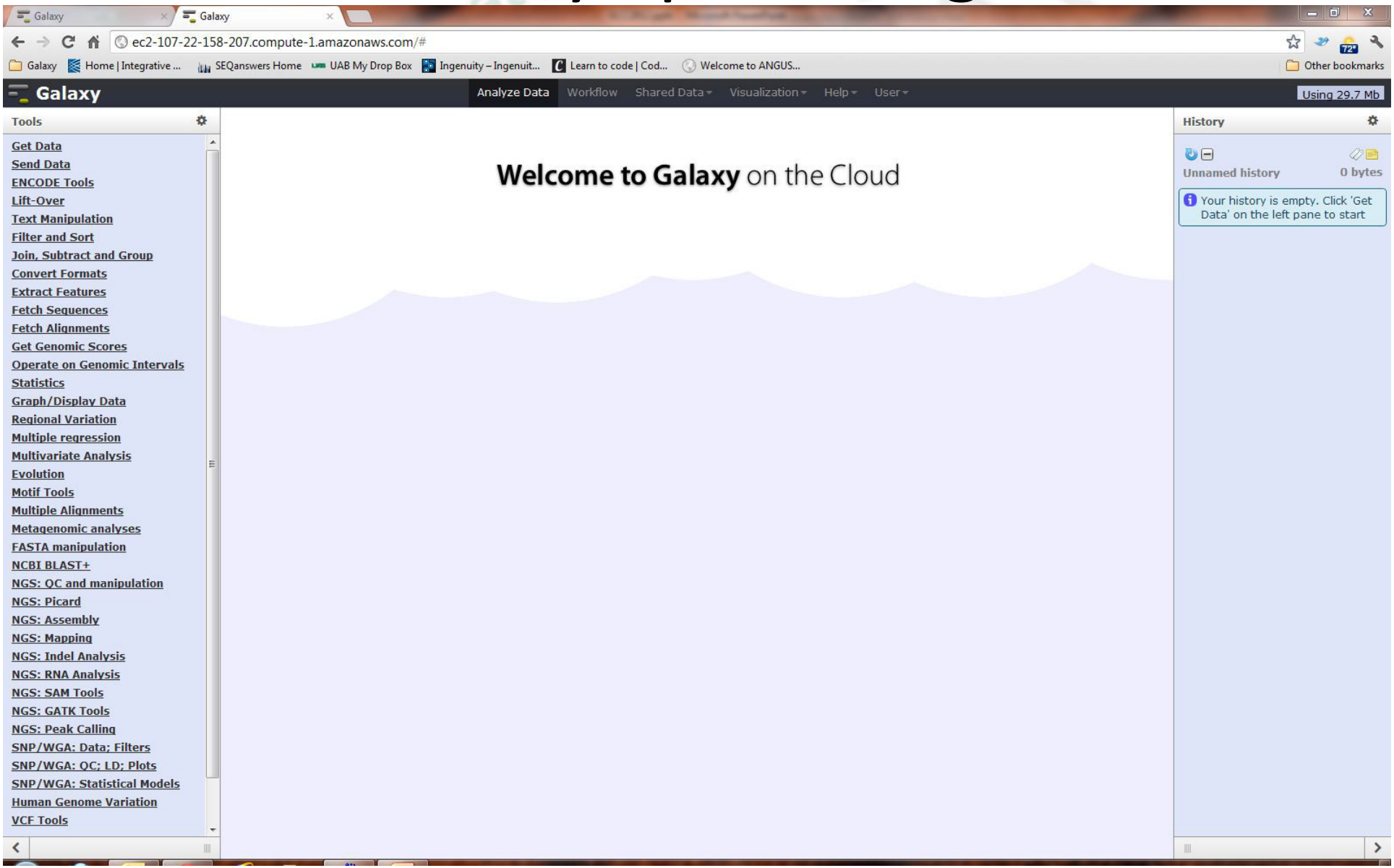
WS9: RNA-Seq Analysis with Galaxy (non-model organism)

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

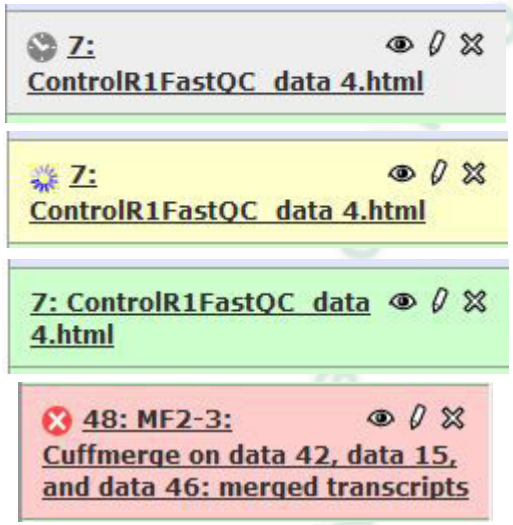
GCC2012
Wednesday, July 25, 2012

Galaxy Splash Page



Random Galaxy icons/colors

Colors



Queued

Running

Completed

Failed

Download/Save

Icons

Display data in browser

Edit attributes

Delete

Edit dataset annotation

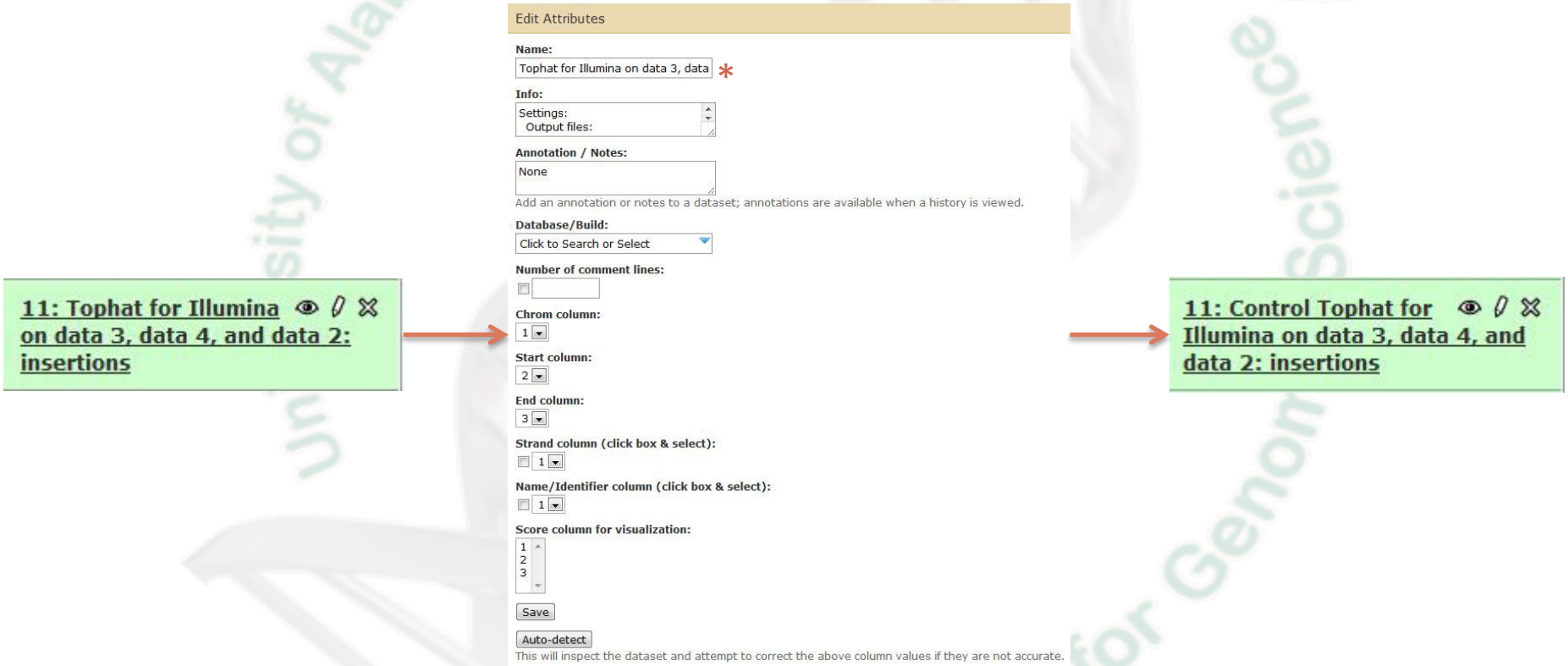
View details

Run this job again

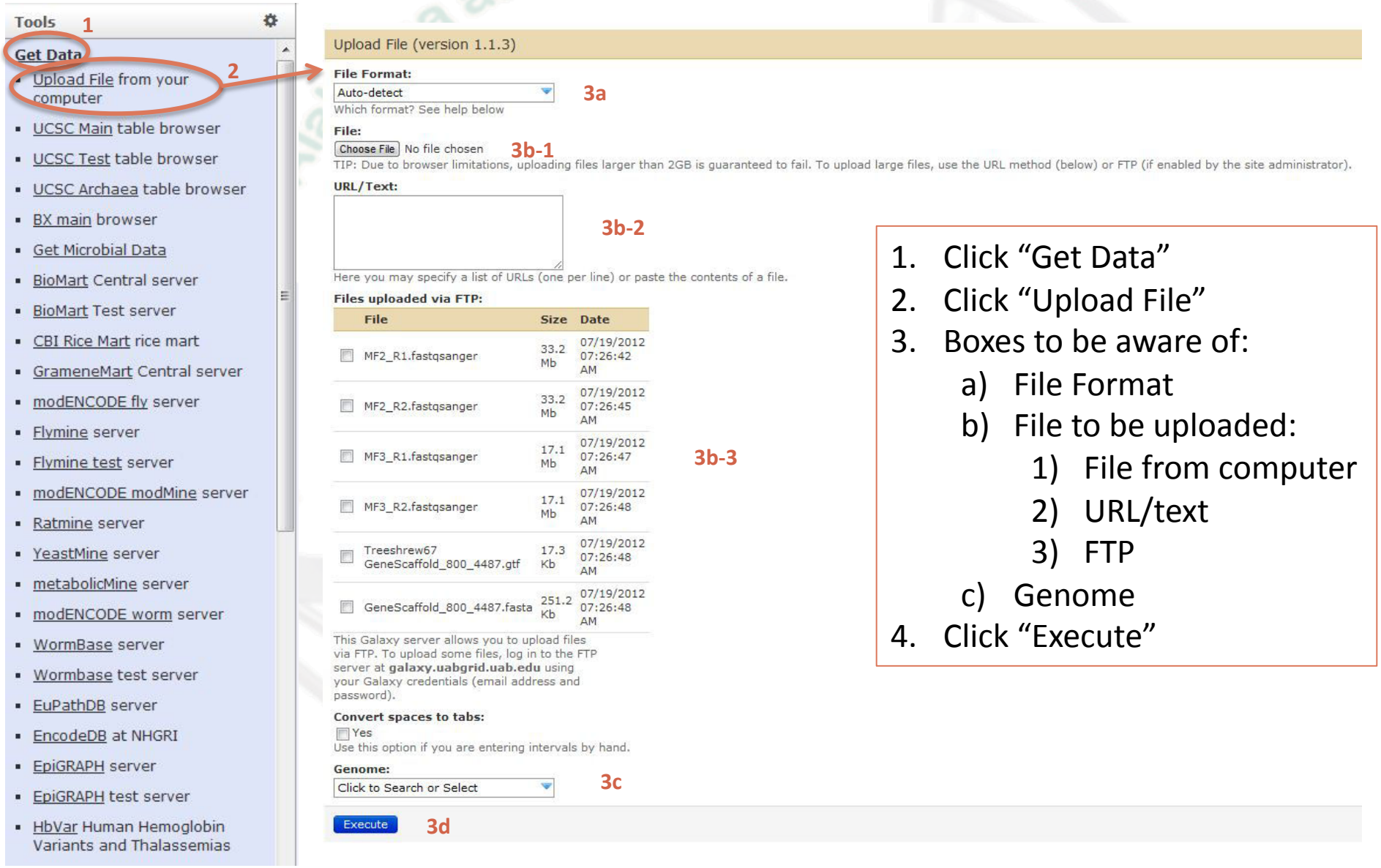
View in Trackster

Edit dataset tags

Edit files in History

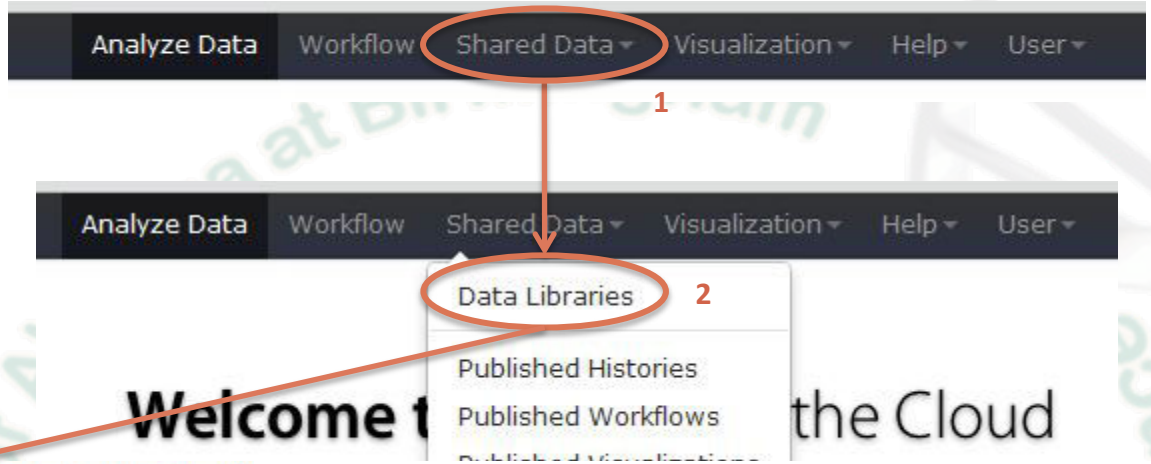


Upload/Import Data



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

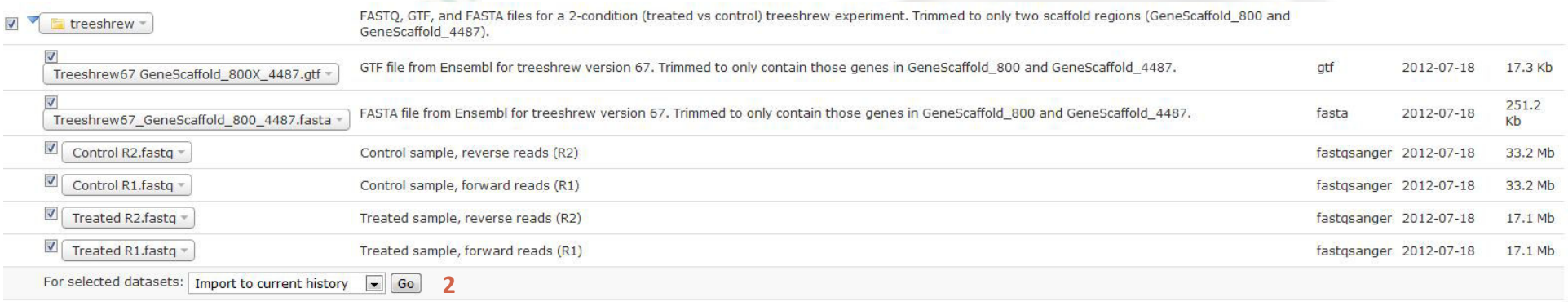
Name	Message	Data type	Date uploaded	File size
mouse	FASTQ and .GTF files for a 2-condition (drugged X control) mouse experiment. Trimmed to just the genes Plekhh2-PigF [chr17:84911234..87424741]. Mouse data kindly provided by Kim Keeiling, PhD at University of Alabama at Birmingham (http://www.microbio.uab.edu/faculty/keeling/index.html)			
Galaxy37-[mm9_RefGene_patched3.gtf].gtf	This dataset is uploading	None	2012-07-18	0 bytes
drugged_mm9_chr17_Plekhh2-PigF_reverse.fastq	Drugged sample, reverse reads	fastqsanger	2012-07-12	1.4 Mb
drugged_mm9_chr17_Plekhh2-PigF_forward.fastq	Drugged sample, forward reads	fastqsanger	2012-07-12	1.4 Mb
control_mm9_chr17_Plekhh2-PigF_forward.fastq	Control sample, forward reads	fastqsanger	2012-07-12	1.5 Mb
control_mm9_chr17_Plekhh2-PigF_reverse.fastq	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).			
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
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
Control R2.fastq	Control sample, reverse reads (R2)	fastqsanger	2012-07-18	33.2 Mb
Control R1.fastq	Control sample, forward reads (R1)	fastqsanger	2012-07-18	33.2 Mb
Treated R2.fastq	Treated sample, reverse reads (R2)	fastqsanger	2012-07-18	17.1 Mb
Treated R1.fastq	Treated sample, forward reads (R1)	fastqsanger	2012-07-18	17.1 Mb

For selected datasets: Import to current history Go

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

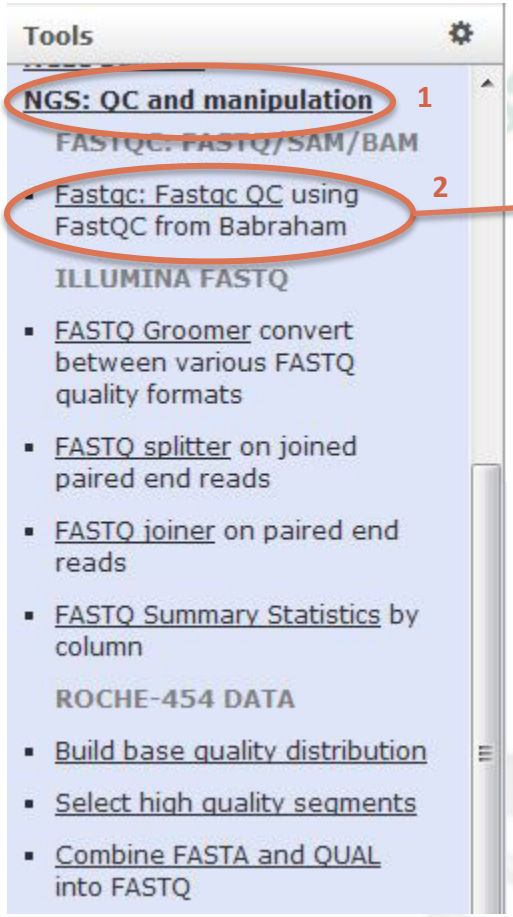
1

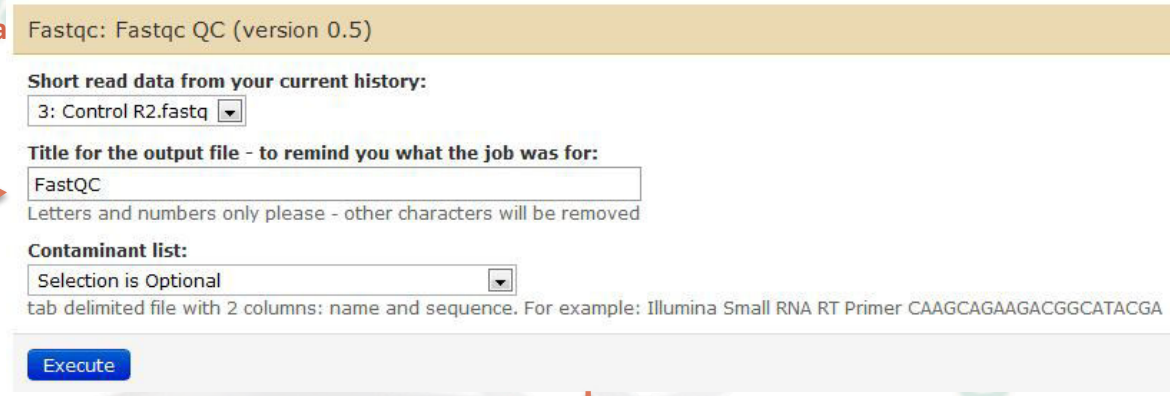


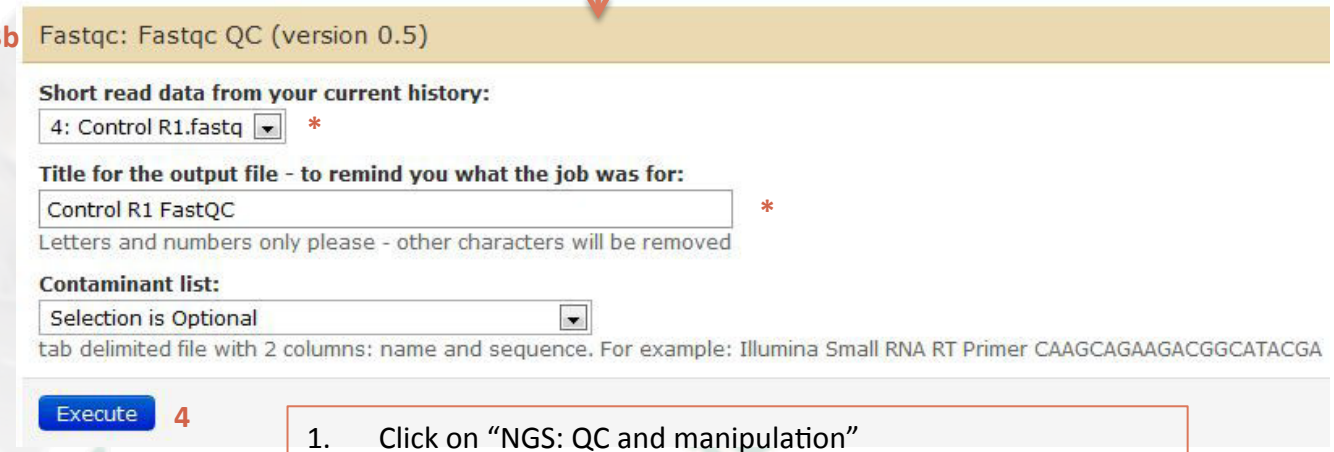
3

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



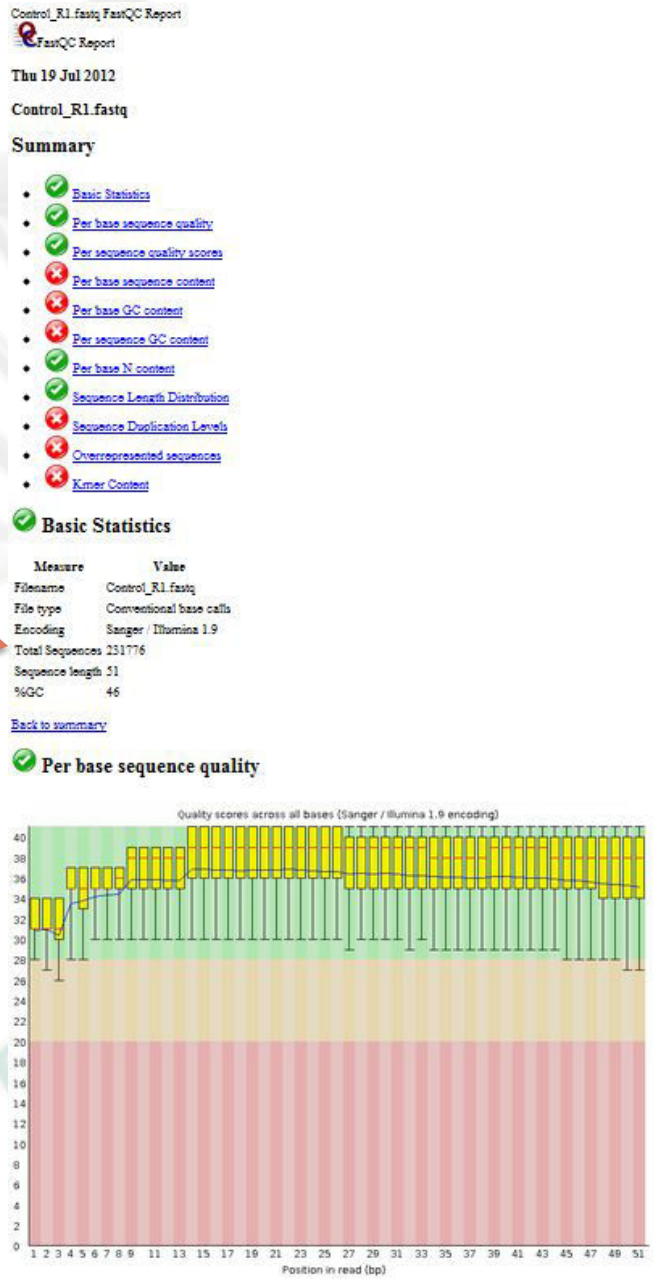




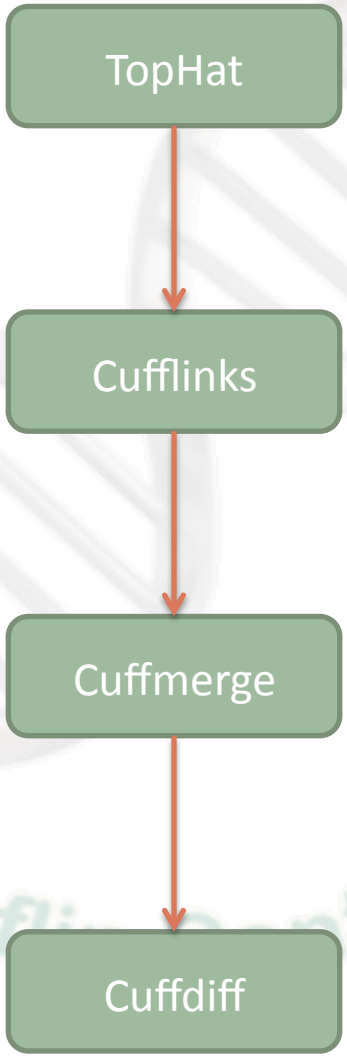
1. Click on “NGS: QC and manipulation”
2. Click on “Fastqc: Fastqc QC
3. Select options:
   a) This is what the window looks like when first opened
   b) Choose fastq file and give it a useful name
4. Click “Execute”
5. Do the exact same thing for the other 3 fastq files

FastQC Output Report

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



RNA-Seq Analysis Pipeline



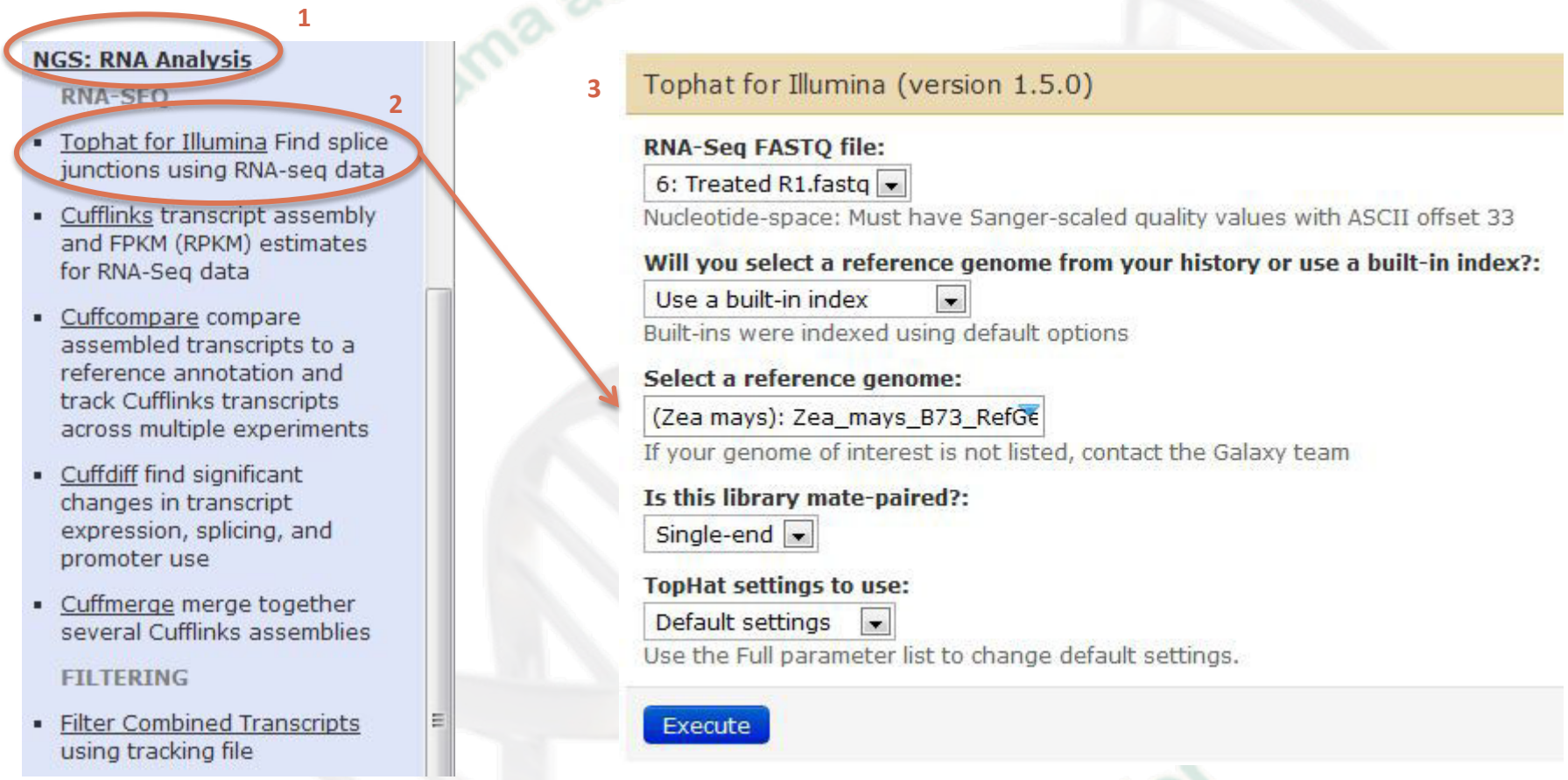
Aligns raw reads to reference genome, identifies splice junctions between exons.

Assembles transcripts, and estimates their abundances.

Merges together several Cufflinks files.

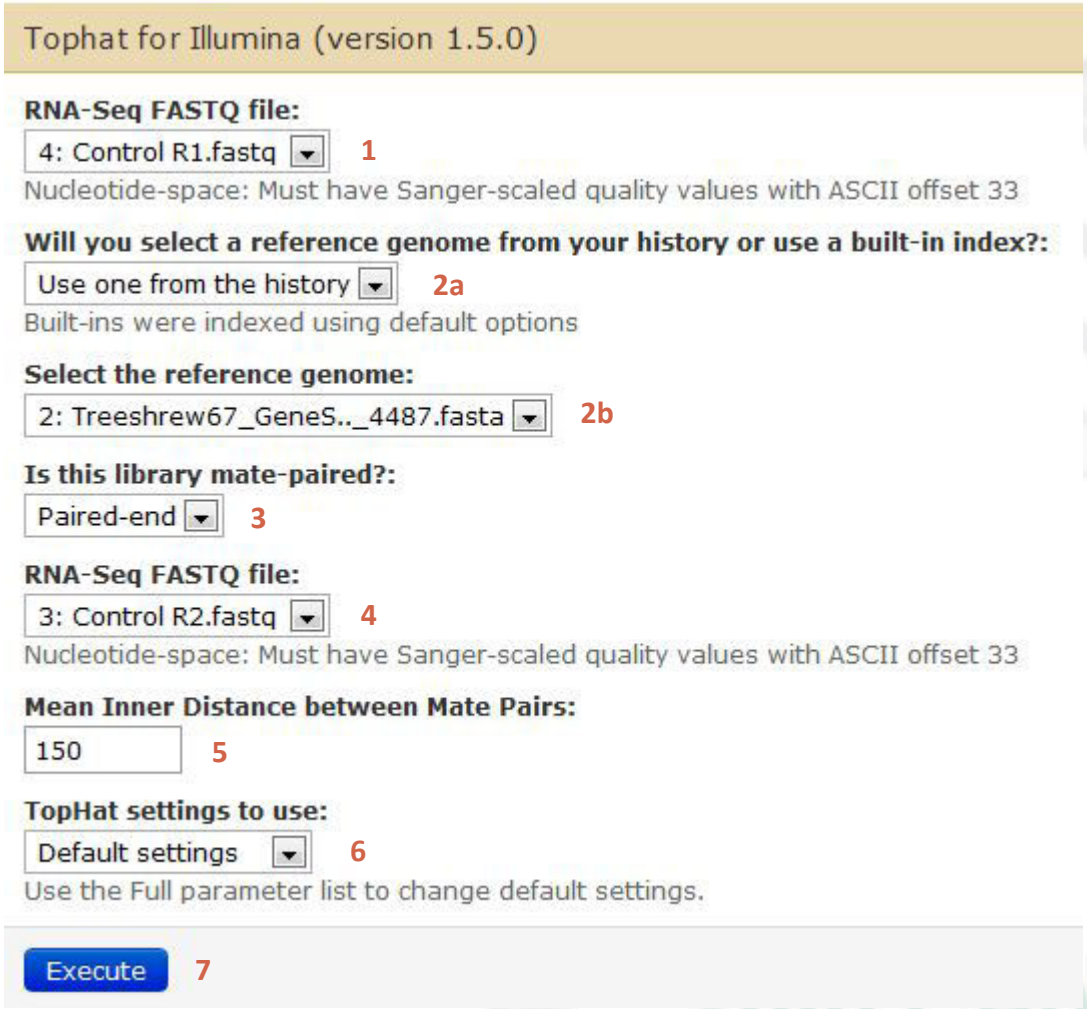
Finds significant changes in transcript expression, splicing, and promoter use.

TopHat



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

TopHat



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:

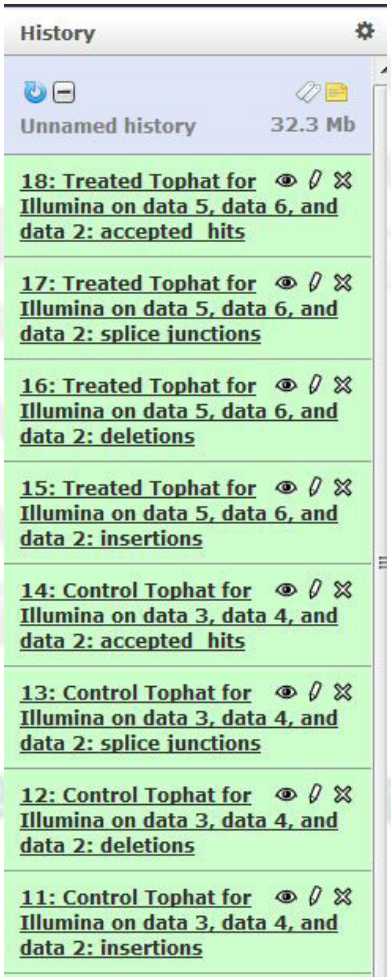
11: Tophat for Illumina on data 3, data 4, and data 2: insertions

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

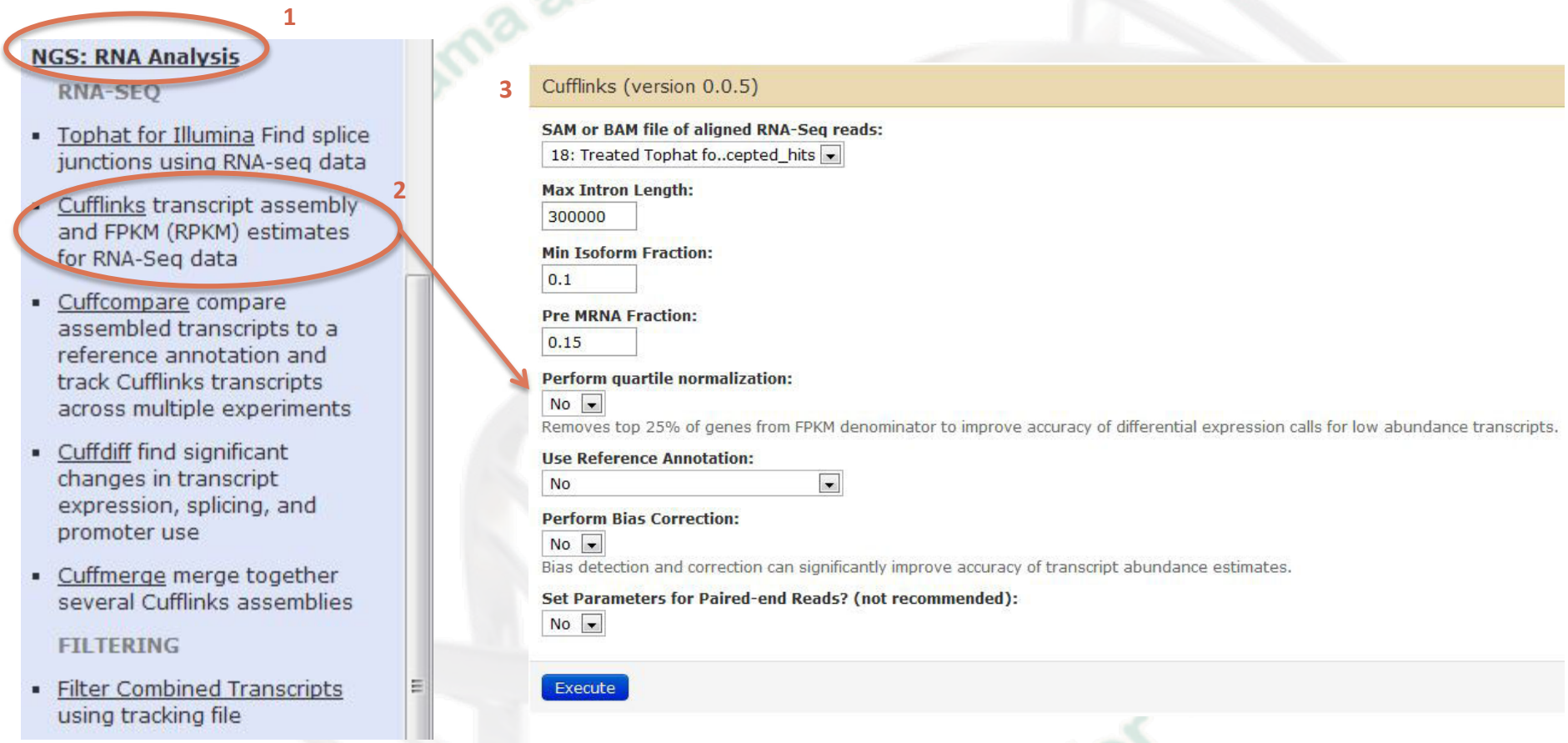
13: Tophat for Illumina on data 3, data 4, and data 2: splice junctions

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.

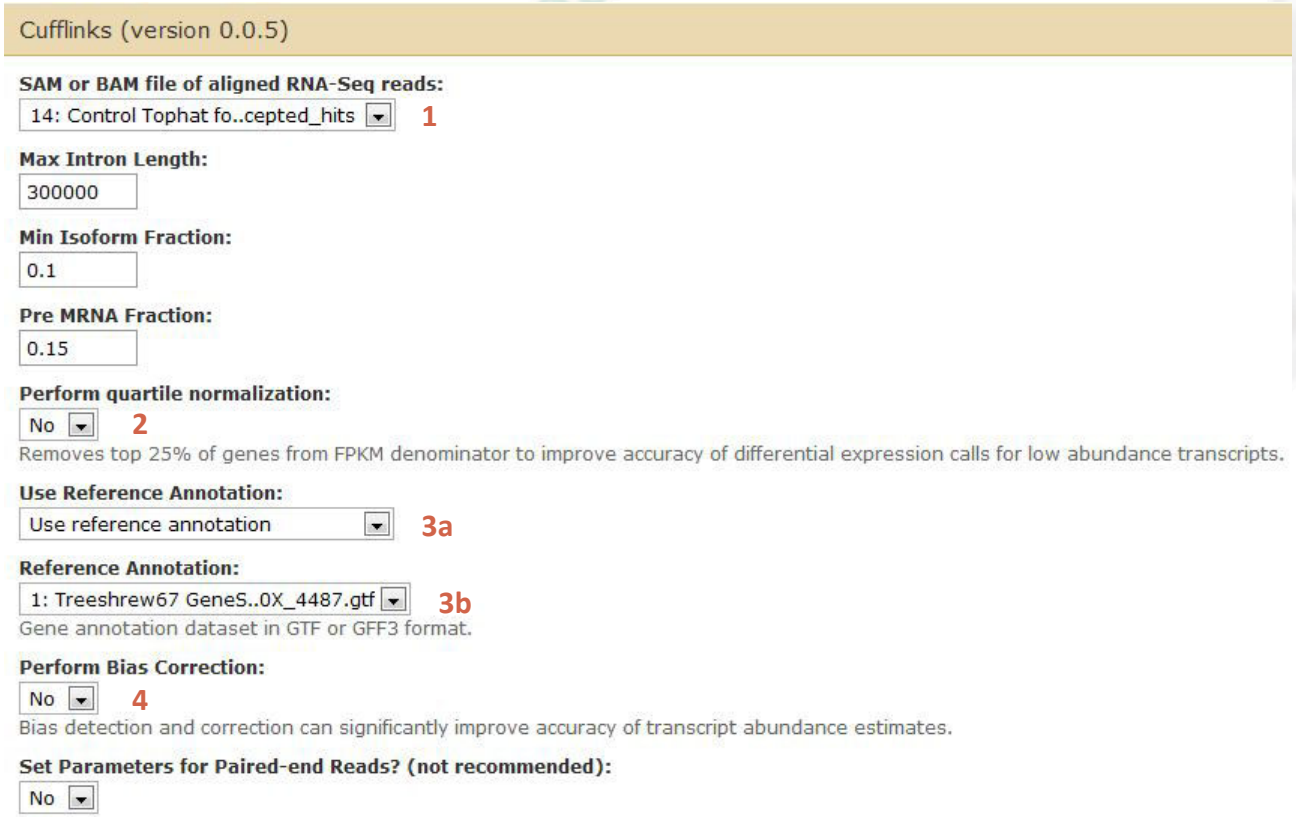


Cufflinks



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

Cufflinks



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:

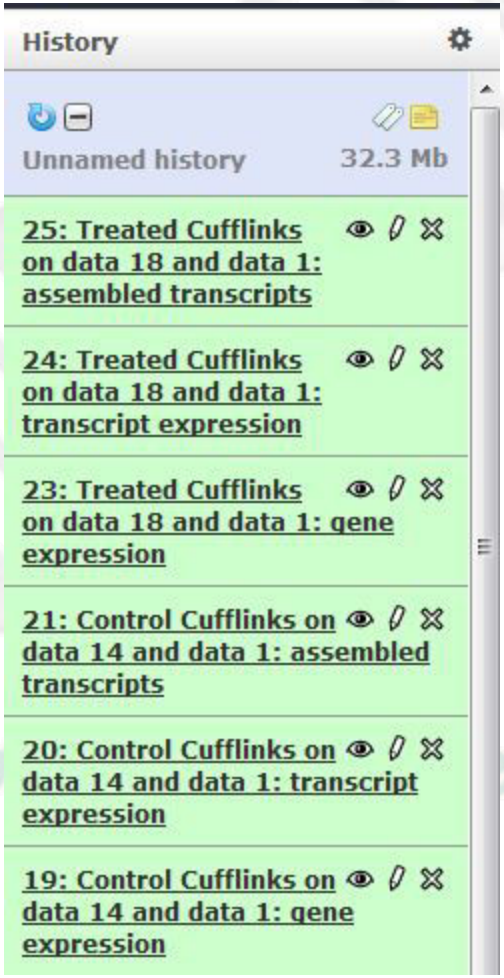
23: Cufflinks on data 18 and data 1: gene expression

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

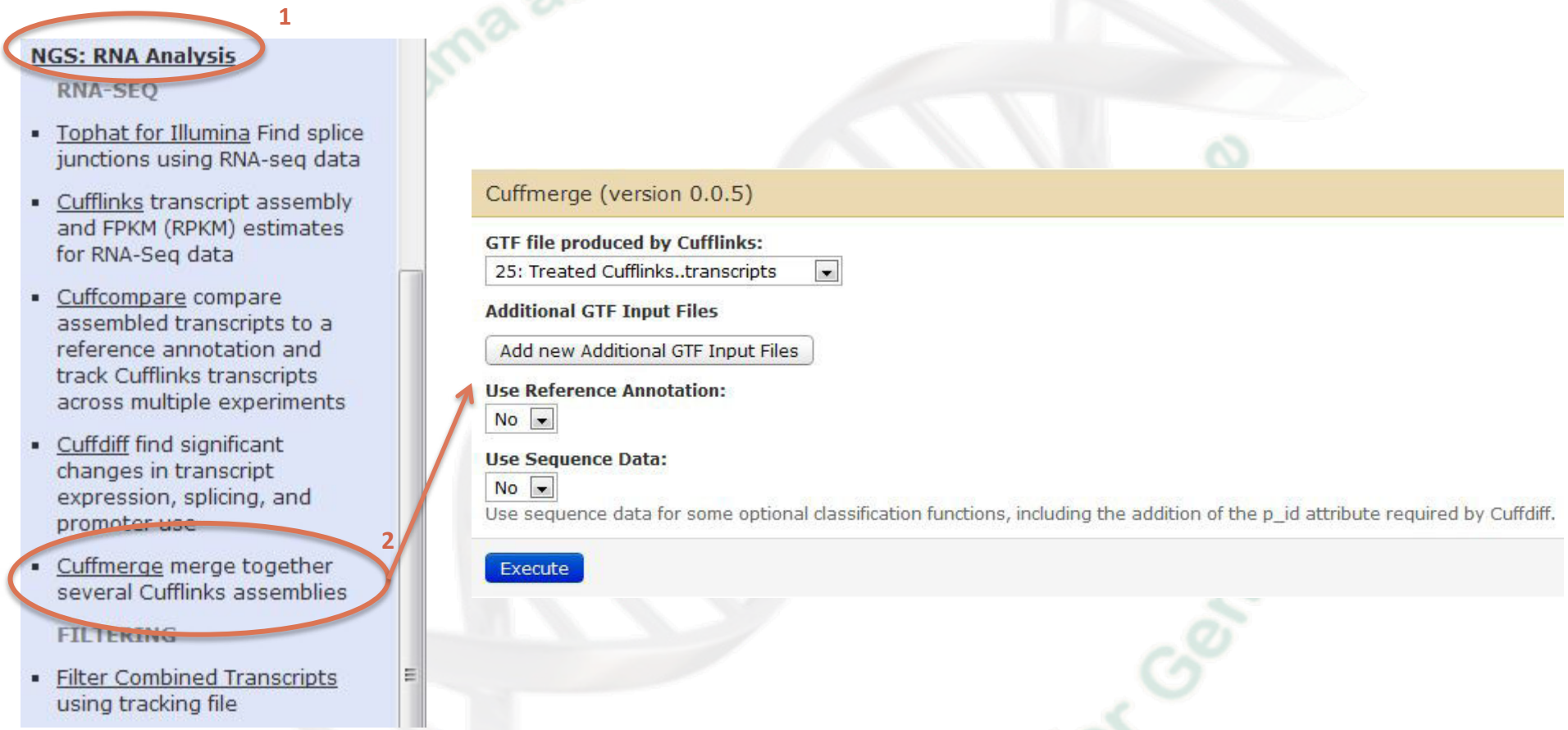
25: Cufflinks on data 18 and data 1: assembled transcripts

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.



Cuffmerge



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

Cuffmerge

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) Slect “Yes” to Use Sequence Data
    b) Choose “History”
    c) Choose FASTA file from history
5. Click “Excecute”

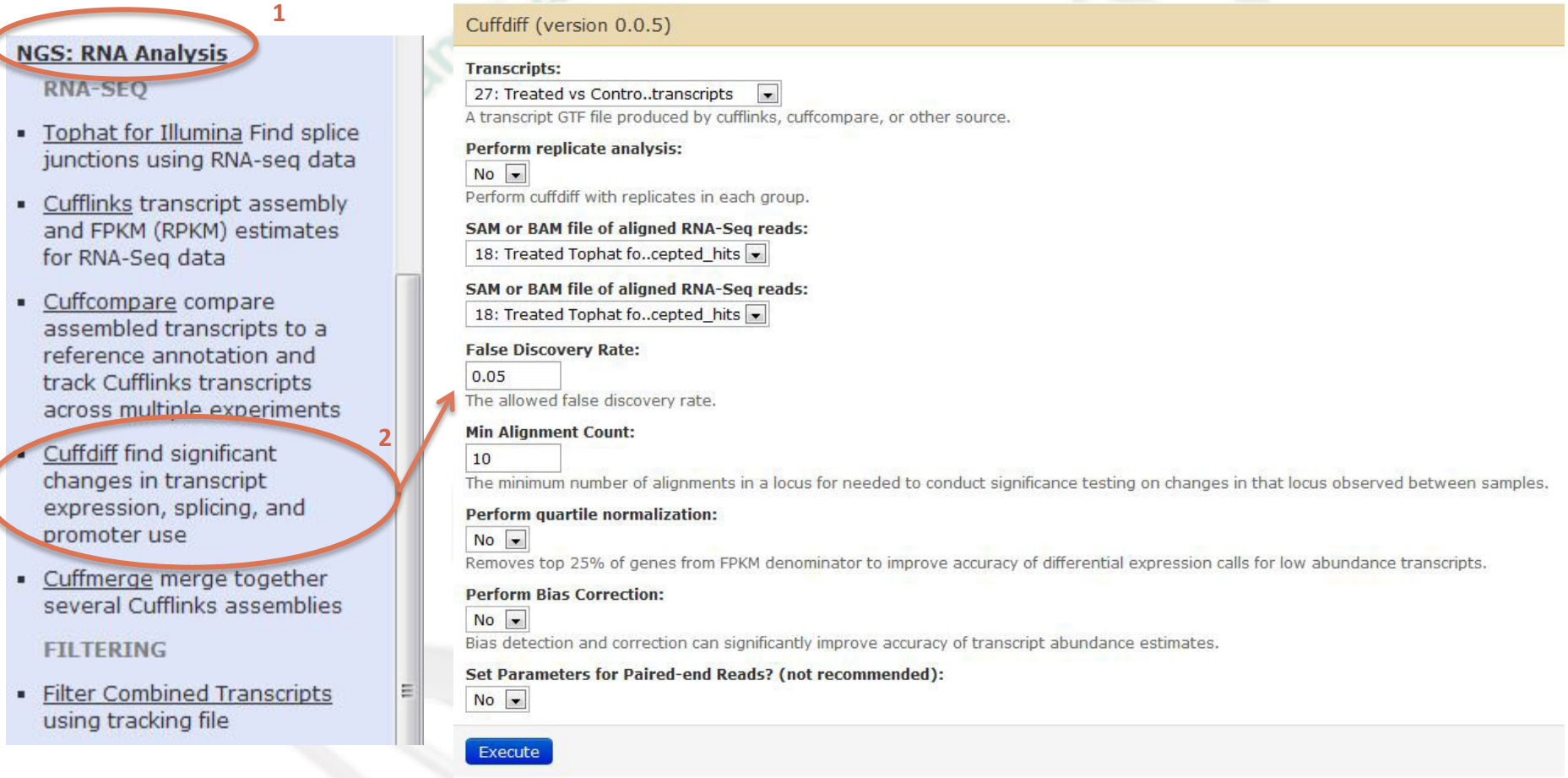
Cuffmerge output files

The following job has been successfully added to the queue:

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

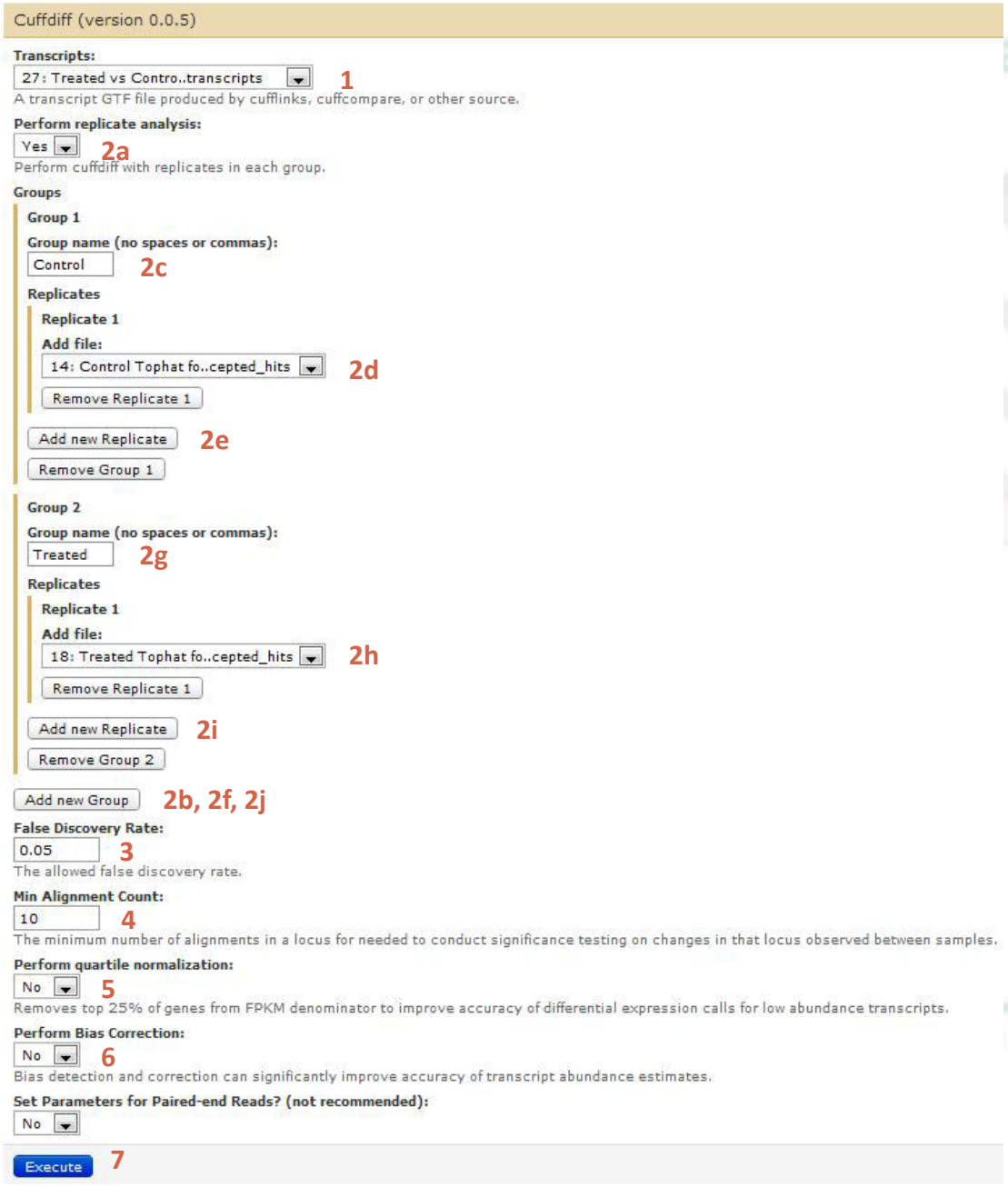
You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Cuffdiff



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

Cuffdiff



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 diffential 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

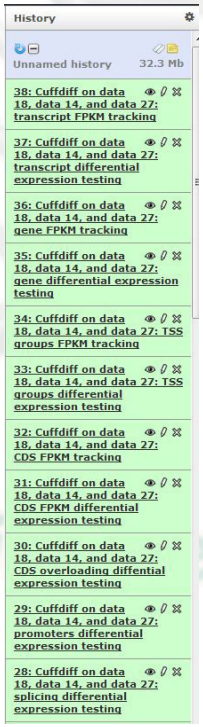
35: Cuffdiff on data 18, data 14, and data 27: gene differential expression testing

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.



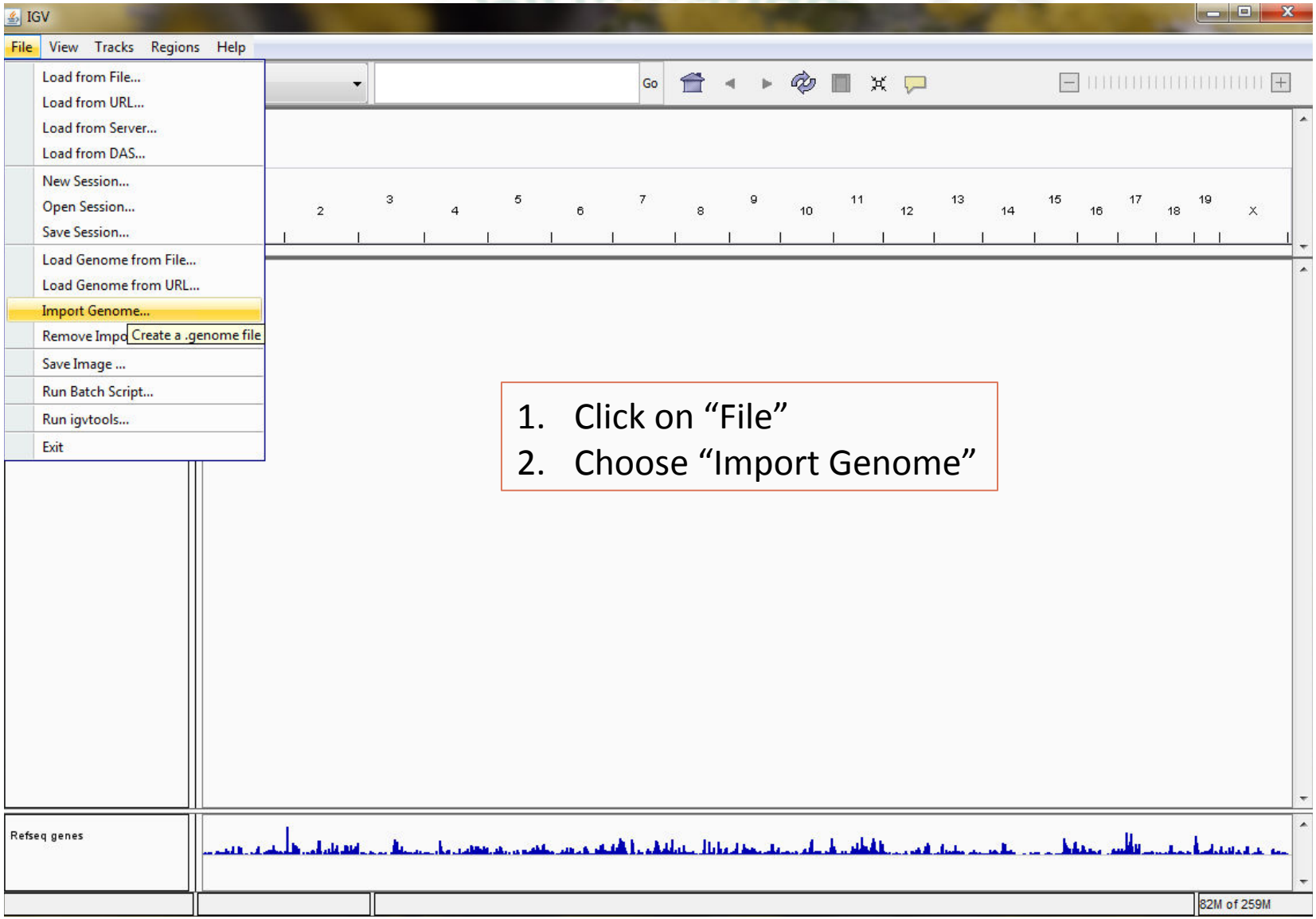
Transcript differential expression testing output

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	si
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

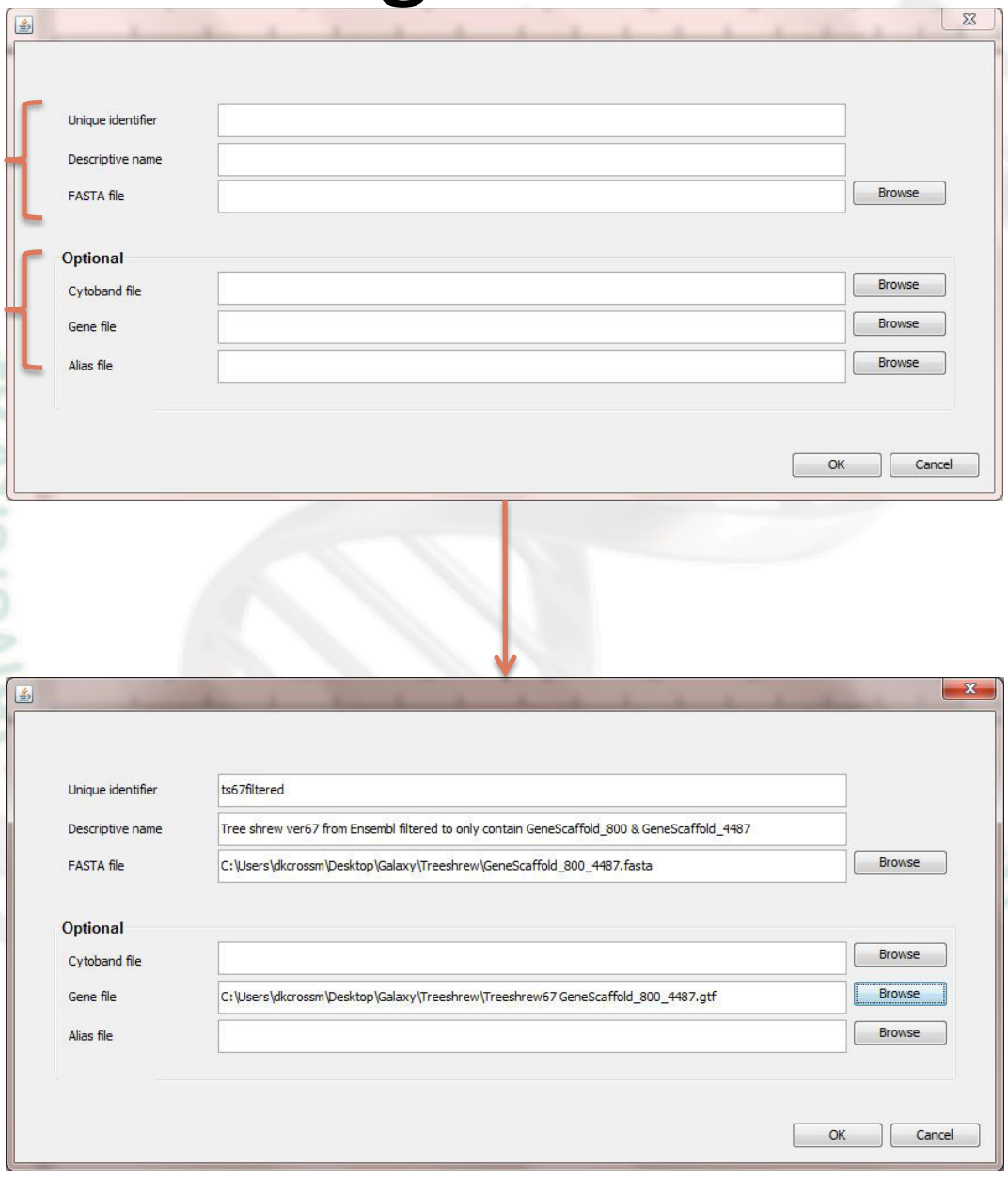


1. Click on “File”
2. Choose “Import Genome”

Create genome in IGV

Required

Optional



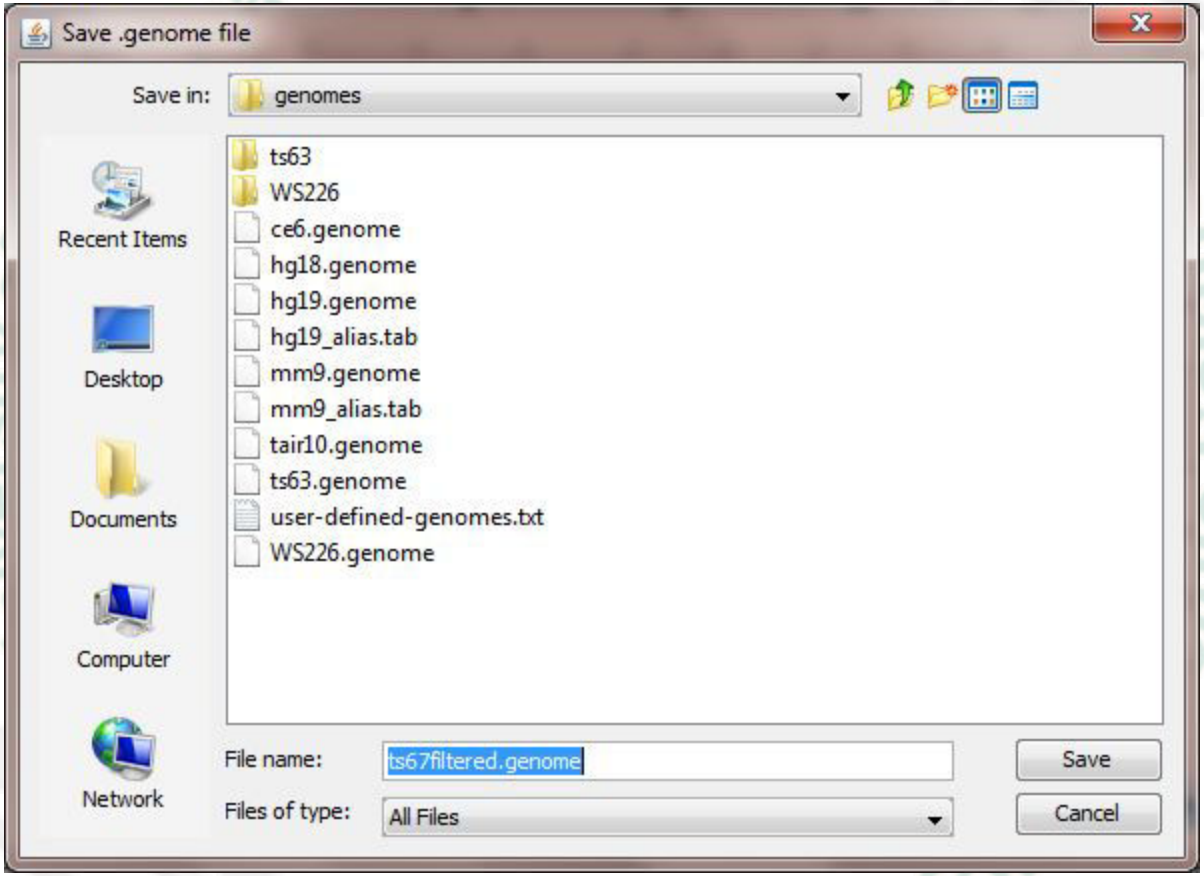
Need:

1. FASTA file (required)

Optional:

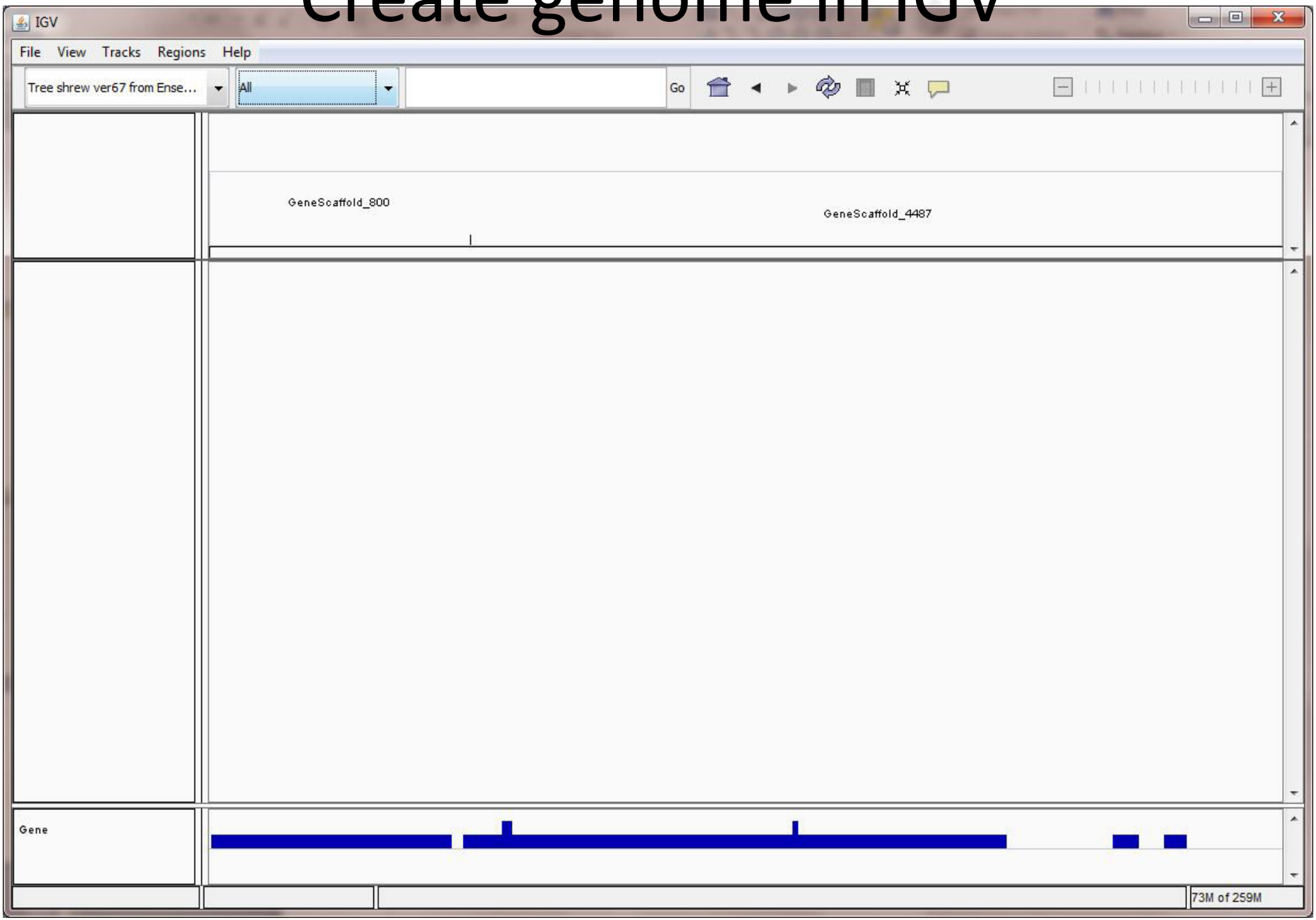
1. Cytoband file
2. Gene file (can use GTF)
3. Alias file

Create genome in IGV

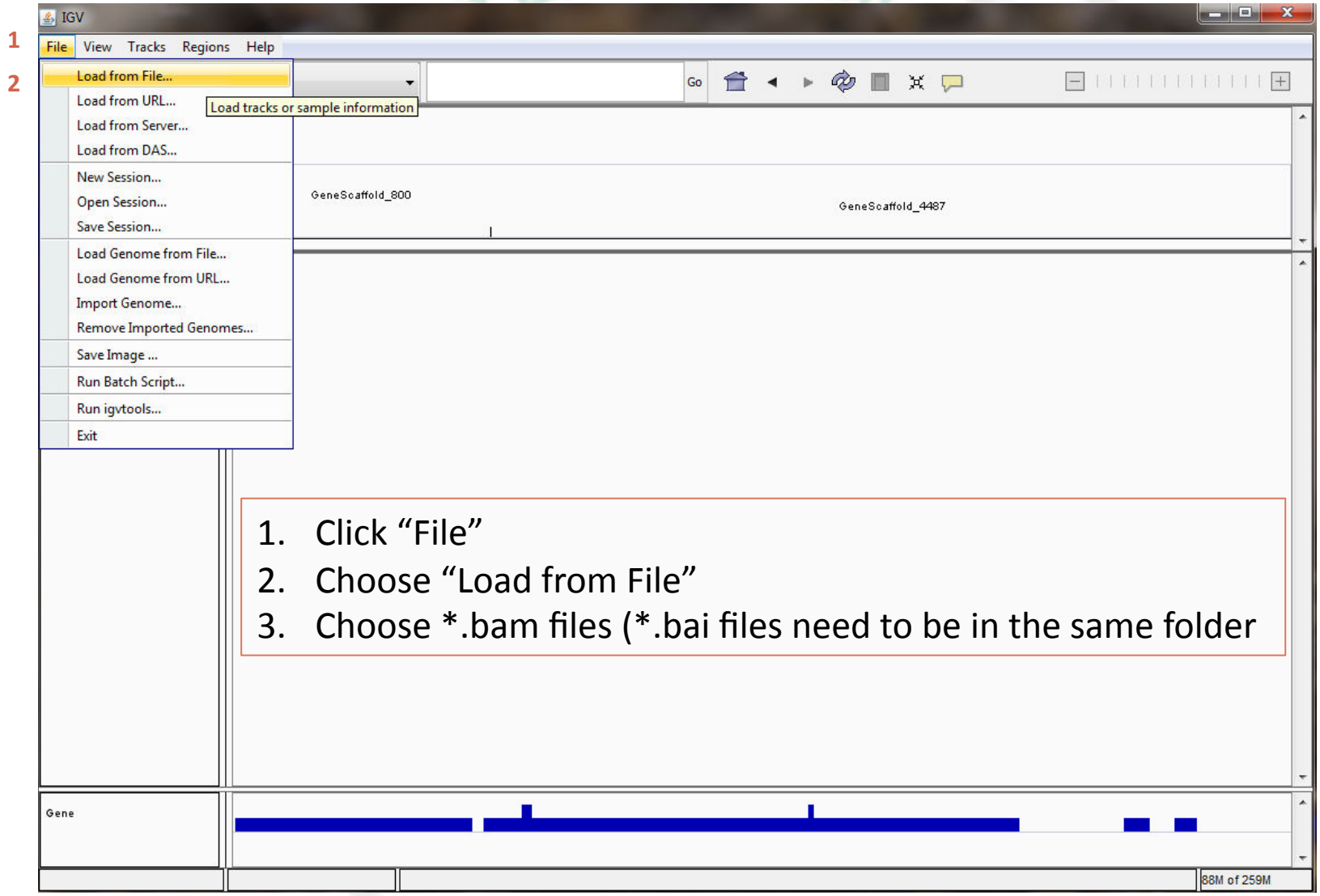


Save the *.genome file

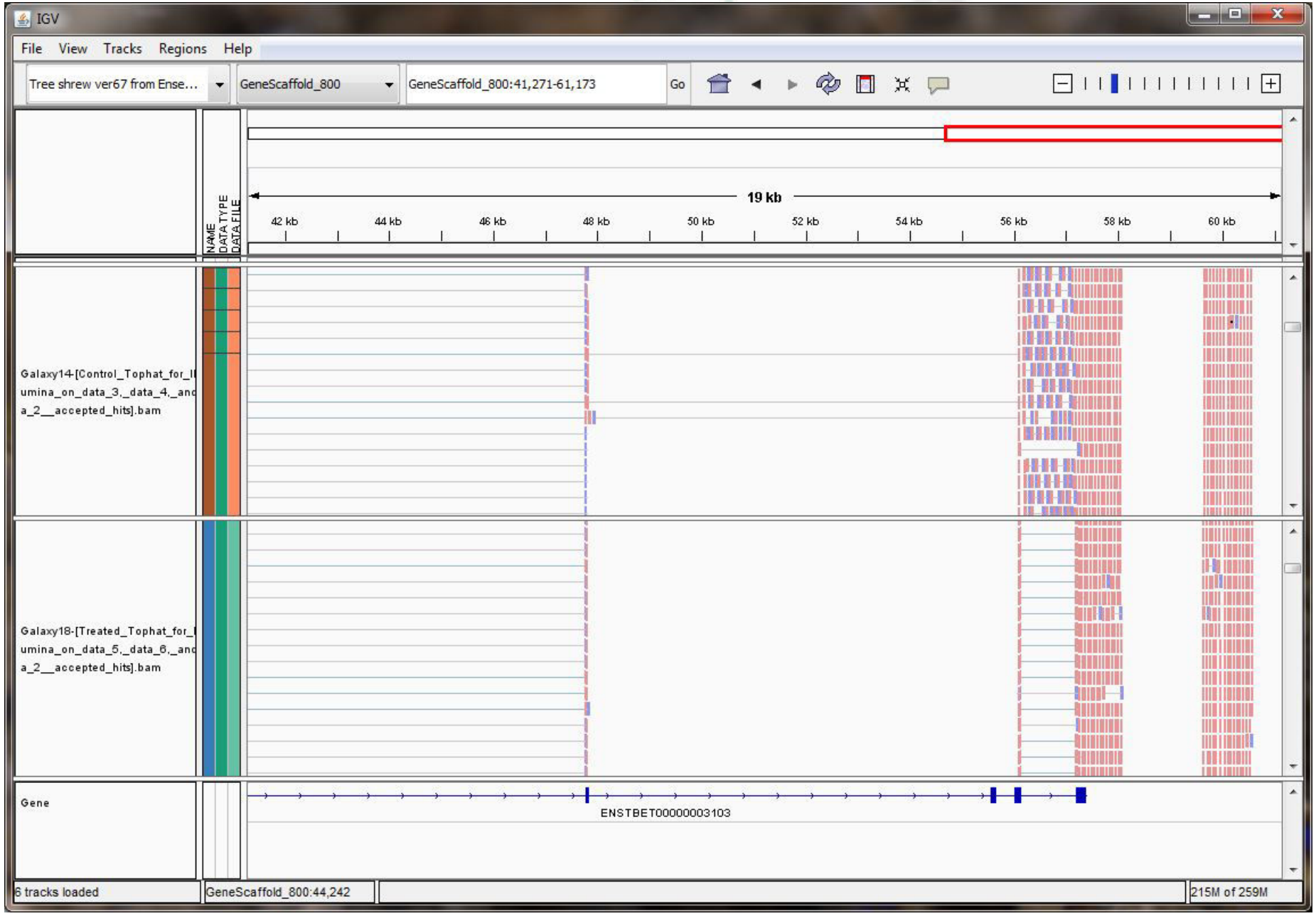
Create genome in IGV



Load aligned BAM files into IGV



IGV


IGV
File View Tracks Regions Help
Tree shrew ver67 from Ense...
GeneScaffold_800
GeneScaffold_800:41,271-61,173
Go
NAME
DATA TYPE
DATA FILE
19 kb
42 kb
44 kb
46 kb
48 kb
50 kb
52 kb
54 kb
56 kb
58 kb
60 kb
Galaxy14-[Control_Tophat_for_I
umina_on_data_3,_data_4,_and
a_2__accepted_hits].bam
Galaxy18-[Treated_Tophat_for_
umina_on_data_5,_data_6,_and
a_2__accepted_hits].bam
Gene
ENSTBET00000003103
6 tracks loaded
GeneScaffold_800:44,242
215M of 259M

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