

# GCC Workshop 9

## RNA-Seq with Galaxy

Curtis Hendrickson ([curtish@uab.edu](mailto:curtish@uab.edu))

David Crossman ([dkcrossm@uab.edu](mailto:dkcrossm@uab.edu))

Jeremy Goecks ([jeremy.goecks@emory.edu](mailto:jeremy.goecks@emory.edu))

# Agenda

- RNA-seq flash review
  - Tophat (RNA-seq read mapping)
  - Cufflinks (Isoform assembly)
  - Cuffmerge (cross-sample consensus)
  - Cuffdiff (expression analysis)
- Galaxy flash review
- RNA-seq exercises
  - installed genome: mouse (mm9)
  - Non-standard genome : tree shrew from Ensembl

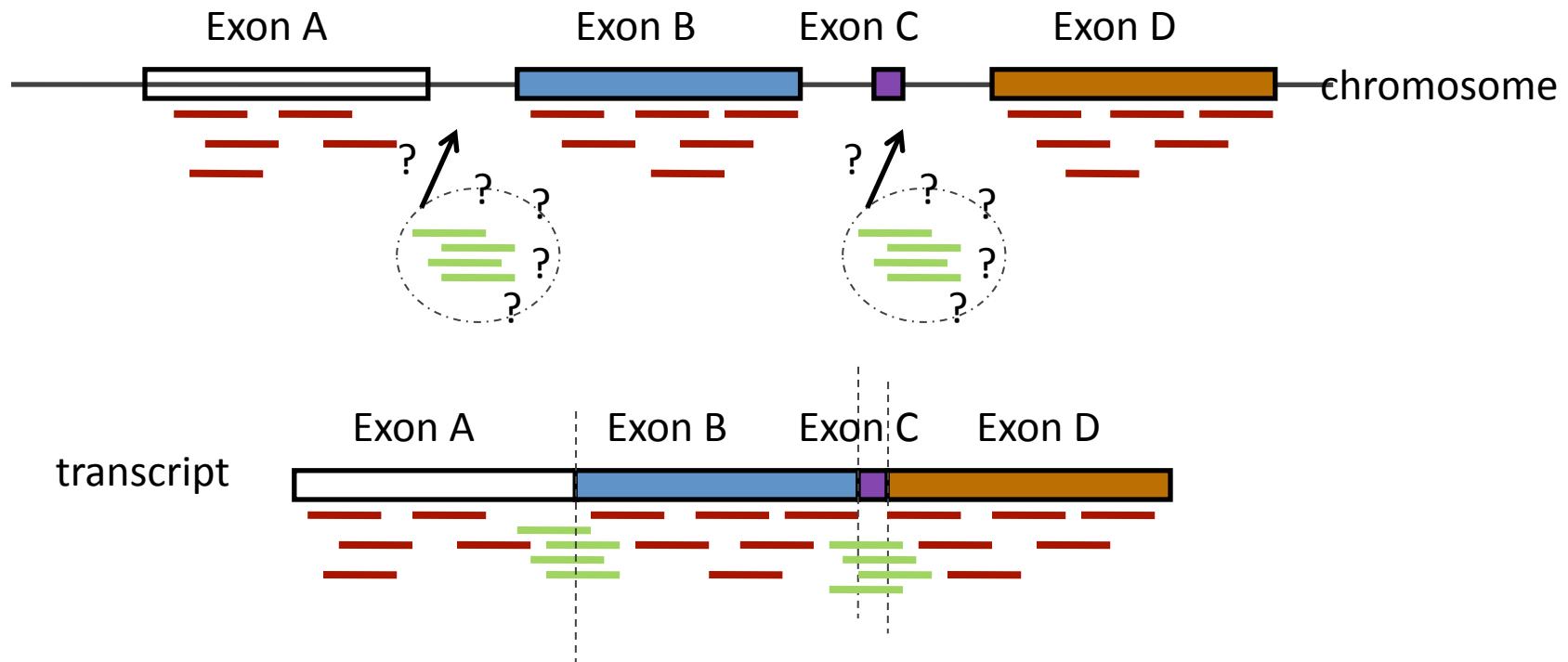
# Why RNA-Seq?

- *Gene/transcript expression quantification*
- Isoform analysis (alternate splicing)
- 5' and 3' UTR analysis
- Sequencing of new genomes (paired genome/transcriptome sequencing)
- etc

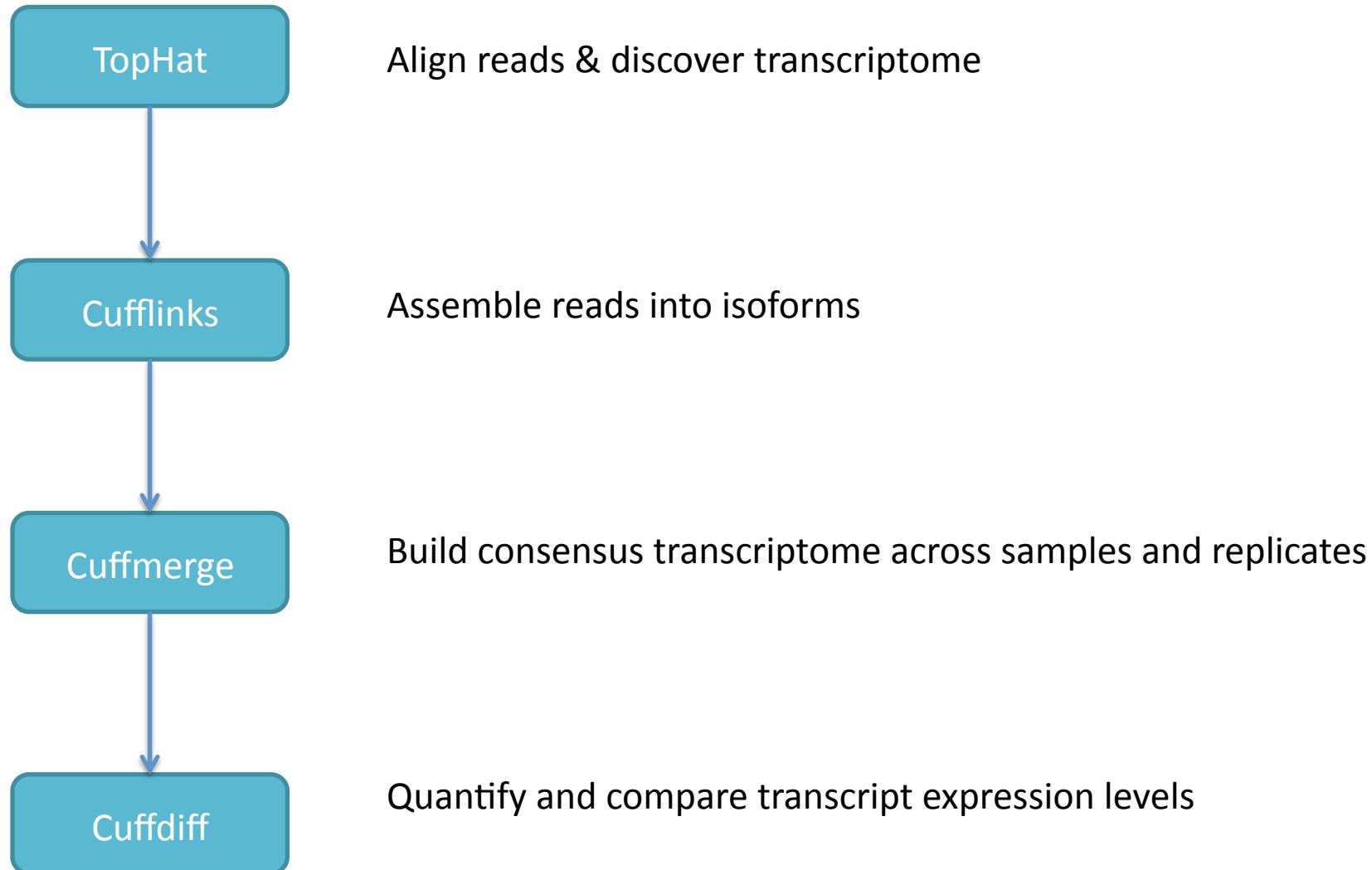
# What type of analysis?

- Compare gene or transcript levels between samples or conditions?
  - Sequence the whole transcriptome
- Discovery of alternate splice forms?
  - $\geq 75\text{bp}$  Paired End for splice site detection
  - Consider target enrichment
  - Even at 100x coverage, a rare (1:100) splice junction would only average 0.4x coverage of its splice junction – not enough to detect.

# Reads in RNA-seq



# An RNA-Seq Analysis Pipeline



See Trapnell *et al*, 2012, *Nat Protocols*

---

## PROTOCOL

# Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks

Cole Trapnell<sup>1,2</sup>, Adam Roberts<sup>3</sup>, Loyal Goff<sup>1,2,4</sup>, Geo Pertea<sup>5,6</sup>, Daehwan Kim<sup>5,7</sup>, David R Kelley<sup>1,2</sup>, Harold Pimentel<sup>3</sup>, Steven L Salzberg<sup>5,6</sup>, John L Rinn<sup>1,2</sup> & Lior Pachter<sup>3,8,9</sup>

---

<sup>1</sup>Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA. <sup>2</sup>Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. <sup>3</sup>Department of Computer Science, University of California, Berkeley, California, USA. <sup>4</sup>Computer Science and Artificial Intelligence Lab, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. <sup>5</sup>Department of Medicine, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. <sup>6</sup>Department of Biostatistics, Johns Hopkins University, Baltimore, Maryland, USA. <sup>7</sup>Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA. <sup>8</sup>Department of Mathematics, University of California, Berkeley, California, USA. <sup>9</sup>Department of Molecular and Cell Biology, University of California, Berkeley, California, USA.  
Correspondence should be addressed to C.T. (cole@broadinstitute.org).

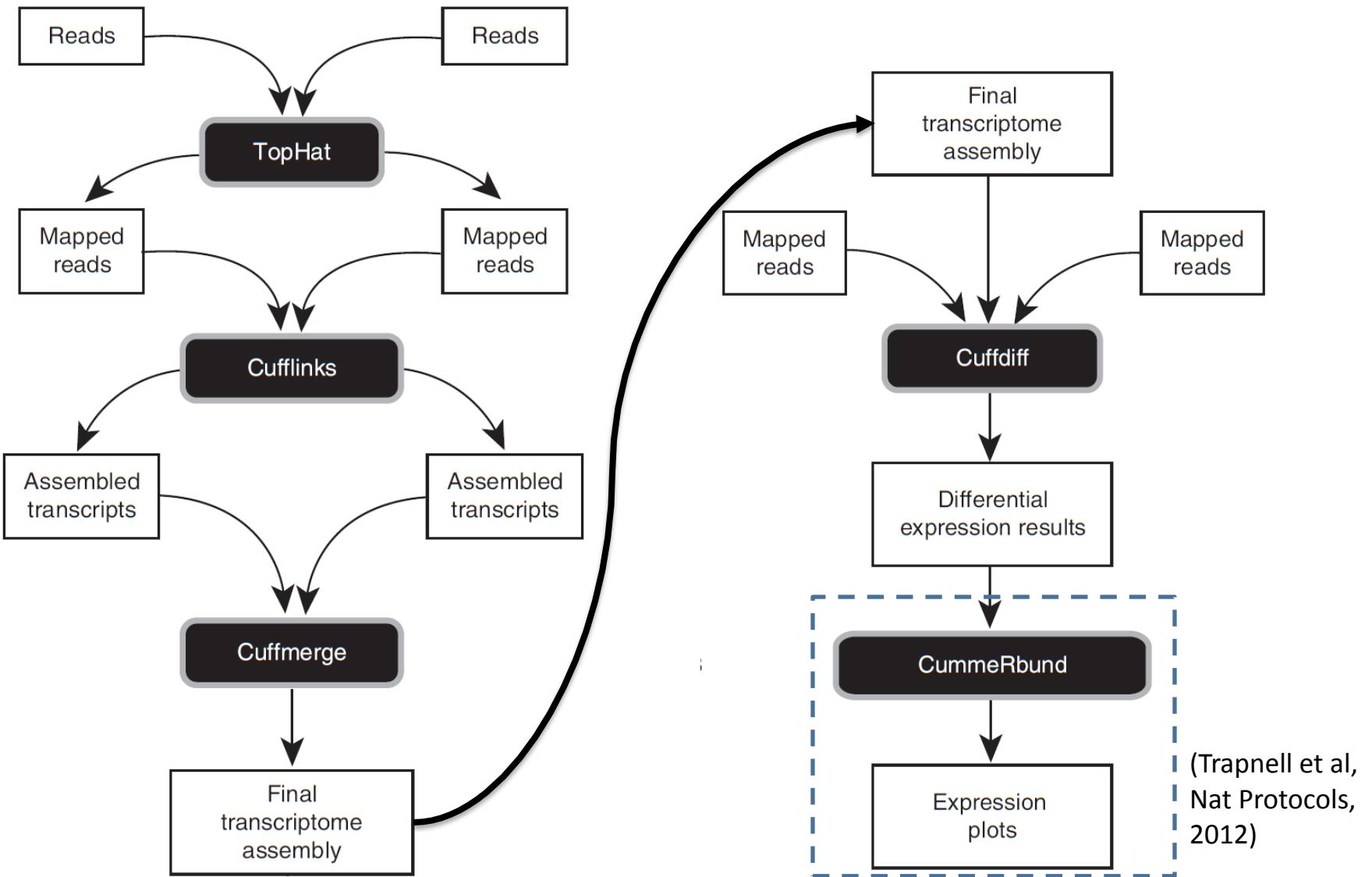
Published online 1 March 2012; doi:10.1038/nprot.2012.016

Recent advances in high-throughput cDNA sequencing (RNA-seq) can reveal new genes and splice variants and quantify expression genome-wide in a single assay. The volume and complexity of data from RNA-seq experiments necessitate scalable, fast and mathematically principled analysis software. TopHat and Cufflinks are free, open-source software tools for gene discovery and comprehensive expression analysis of high-throughput mRNA sequencing (RNA-seq) data. Together, they allow biologists to identify new genes and new splice variants of known ones, as well as compare gene and transcript expression under two or more conditions. This protocol describes in detail how to use TopHat and Cufflinks to perform such analyses. It also covers several accessory tools and utilities that aid in managing data, including CummeRbund, a tool for visualizing RNA-seq analysis results. Although the procedure assumes basic informatics skills, these tools assume little to no background with RNA-seq analysis and are meant for novices and experts alike. The protocol begins with raw sequencing reads and produces a transcriptome assembly, lists of differentially expressed and regulated genes and transcripts, and publication-quality visualizations of analysis results. The protocol's execution time depends on the volume of transcriptome sequencing data and available computing resources but takes less than 1 d of computer time for typical experiments and ~1 h of hands-on time.

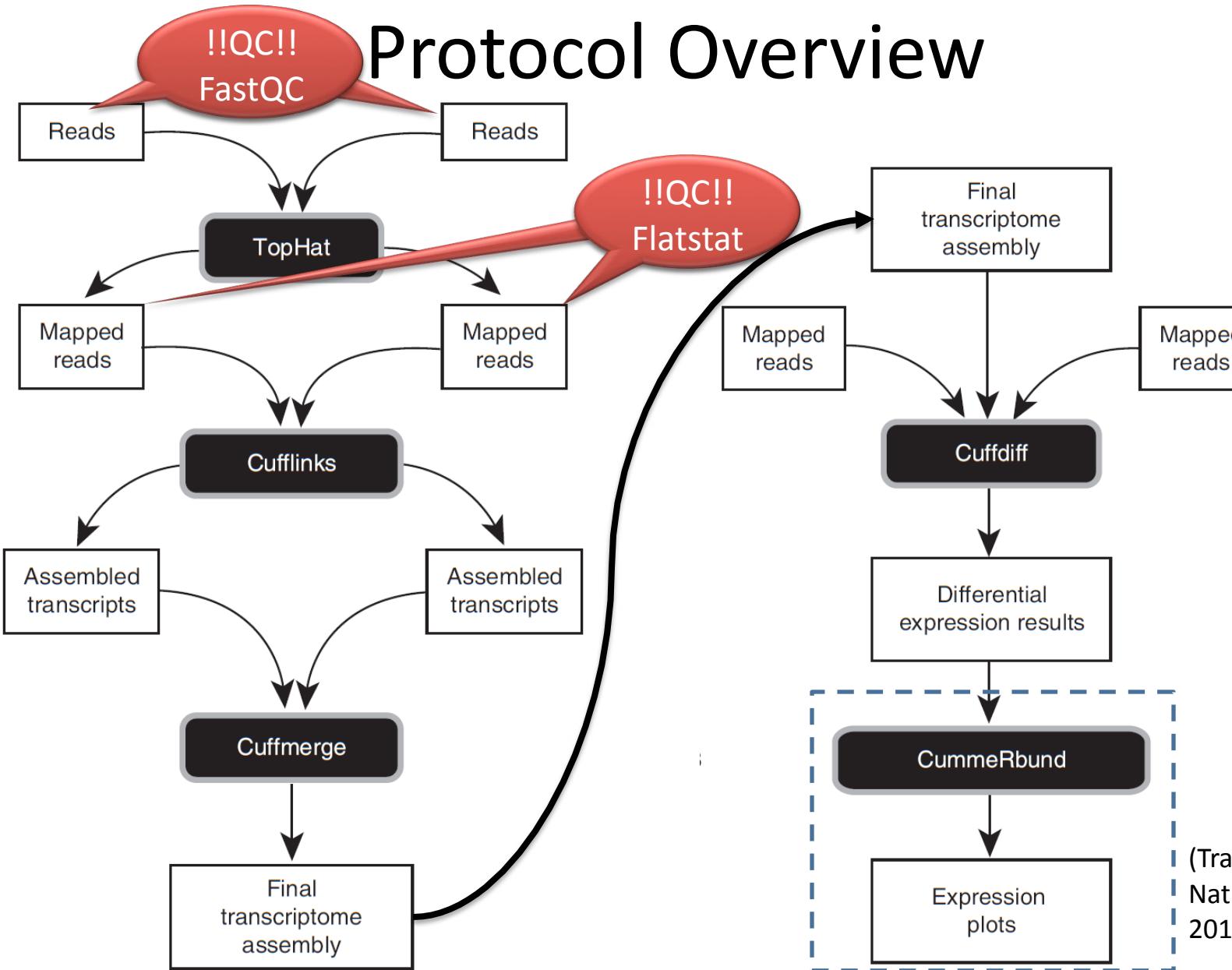
# Types of data

- FASTA <1G Genomes
- FASTQ 1-12G/ea Sequencing reads
- BAM 1-5G/ea Mapped Reads (*binary*)
- BED <1G Genomic intervals
- GTF < 1G Gene/transcript location

# 2-Condition Differential Expression Protocol Overview



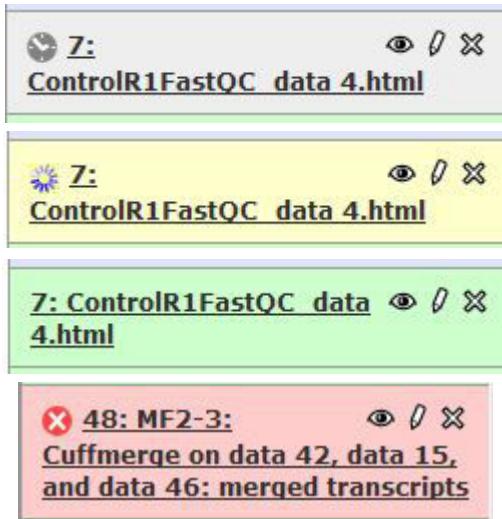
# 2-Condition Differential Expression Protocol Overview



(Trapnell et al,  
Nat Protocols,  
2012)

# Random Galaxy icons/colors

## Colors



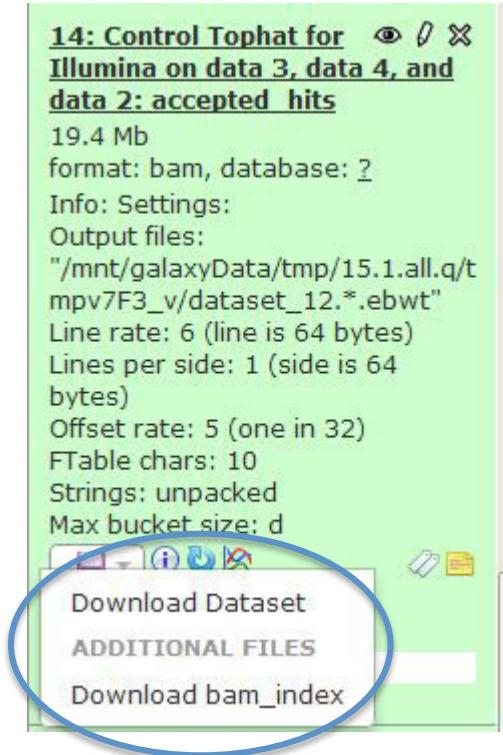
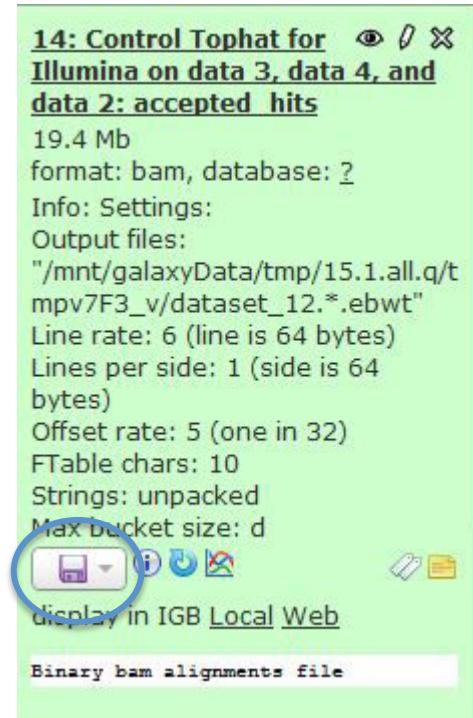
Queued

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

# In the beginning there were reads...

*Get them from Shared Data*

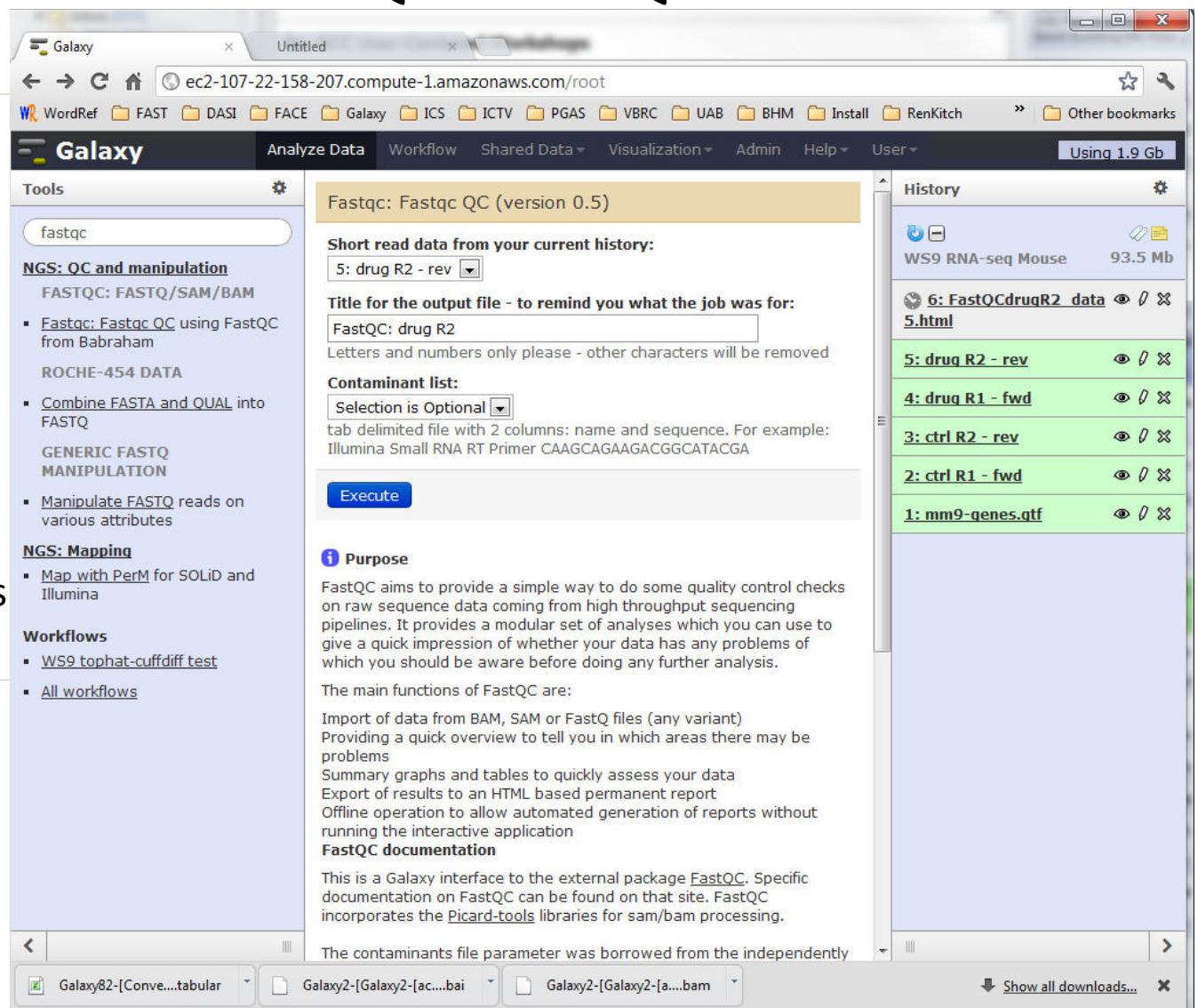
The screenshot shows the Galaxy web interface. At the top, there is a toolbar with tabs: Analyze Data, Workflow, Shared Data (which is highlighted with a blue oval and has a blue arrow pointing down to it), Visualization, Help, and User. Below the toolbar, the main content area has a "Welcome to the Cloud" message. On the left, there is a sidebar titled "Data Library 'WS9: RNA-Seq Analysis with Galaxy'". This sidebar lists several items under categories like "mouse", "gene annotation", "reads", and "treeshrew". Each item has a small icon and a brief description. To the right of the sidebar is a table listing datasets. The table has columns: Name, Message, Data type, Date uploaded, and File size. The table contains entries for various FASTQ, GTF, and Fasta files from the "mouse", "gene annotation", "reads", and "treeshrew" categories. A blue arrow labeled "3" points from the "Data Library" title in the sidebar down to the "Data Library 'WS9: RNA-Seq Analysis with Galaxy'" title in the main content area.

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)

Name	Message	Data type	Date uploaded	File size
mouse	FASTQ and .GTF files for a 2-condition (drug vs ctrl) mouse experiment.			
gene annotation	iGenome/Mus_musculus/UCSC/mm9/genes.gtf See <a href="http://cufflinks.cbcb.umd.edu/igenomes.html">http://cufflinks.cbcb.umd.edu/igenomes.html</a> Specially prepared to work well with cufflinks/diff/compare.	gtf	2012-07-20	80.3 Mb
reads	Reads trimmed to just mm9 chr8: 109572500-110112500 (Cog8-Psmd7). Mouse data kindly provided by Kim Keeeling, PhD at University of Alabama at Birmingham ( <a href="http://www.microbio.uab.edu/faculty/keeling/index.html">http://www.microbio.uab.edu/faculty/keeling/index.html</a> )	fastqsanger	2012-07-23	3.3 Mb
ctrl R1 (fwd)	None	fastqsanger	2012-07-23	3.3 Mb
ctrl R2 (rev)	None	fastqsanger	2012-07-23	3.3 Mb
drug R1 (fwd)	None	fastqsanger	2012-07-23	3.3 Mb
drug R2 (rev)	None	fastqsanger	2012-07-23	3.3 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

# So we used FASTQC to QC them...

- Tools >  > Tool Search
- Enter “fastqc”
- Select the tool
- Enter a data set & Execute
- Trick for quick entry of multiple samples:
  - Immediately hit “BACK”
  - Update parameters and re-execute

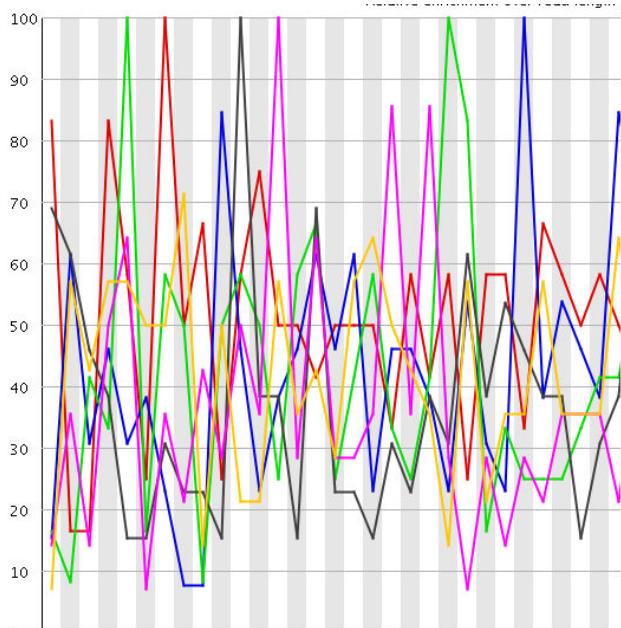


The screenshot shows the Galaxy web interface with the following details:

- Header:** Galaxy, Untitled, ec2-107-22-158-207.compute-1.amazonaws.com/root.
- Tool Search:** The search bar contains "fastqc".
- Tool Details:** The "Fastqc: Fastqc QC (version 0.5)" tool is selected. It has the following configuration:
  - Short read data from your current history: 5: drug R2 - rev
  - Title for the output file - to remind you what the job was for: FastQC: drug R2
  - Contaminant list: Selection is Optional (checkbox)
- Execute Button:** A blue "Execute" button is present.
- Purpose Section:** Describes FastQC's aim to provide quality control checks on raw sequence data.
- Functions Section:** Lists functions like Import of data from BAM, SAM or FastQ files, Providing a quick overview, Summary graphs and tables, Export of results to an HTML based permanent report, Offline operation, and FastQC documentation.
- Footnote:** Notes that the contaminants file parameter was borrowed from independently developed software.
- History Panel:** Shows a list of recent jobs:
  - WS9 RNA-seq Mouse (93.5 Mb)
  - 6: FastQCdrugR2\_data (5.html)
  - 5: drug R2 - rev
  - 4: drug R1 - fwd
  - 3: ctrl R2 - rev
  - 2: ctrl R1 - fwd
  - 1: mm9-genes.gtf
- Bottom Navigation:** Galaxy82-[Conve....tabular], Galaxy2-[Galaxy2-[ac....bai], Galaxy2-[Galaxy2-[a...bam], Show all downloads...

# FastQC Output Report

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



Control\_R1.fastq FastQC Report  
FastQC Report  
Thu 19 Jul 2012  
Control\_R1.fastq  
Summary

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.gtf

Basic Statistics

- 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

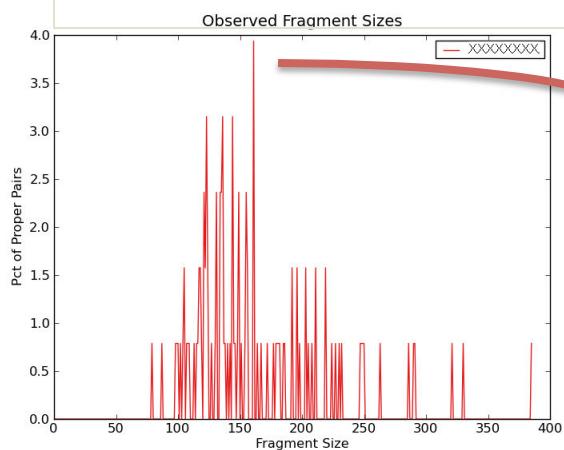
Per base sequence quality

Quality scores across all bases (Sanger / Illumina 1.9 encoding)

# Tophat: map reads, create transcriptome

Paired-end, so 2 FASTQ per Tophat run

- Select forward FASTQ
- Set the ref. genome
- Set PAIRed ends
- Select reverse FASTQ
- Select inner distance (get from sequencing group)



Galaxy

ec2-107-22-158-207.compute-1.amazonaws.com/root#

Galaxy Analyze Data Workflow Shared Data Visualization Admin Help User Using 1.9 Gb

Tools tophat 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 Workflows

**Tophat for Illumina (version 1.5.0)**

RNA-Seq FASTQ file: 2: ctrl R1 - fwd 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: mm9 If your genome of interest is not listed, contact the Galaxy team Is this library mate-paired?: Paired-end RNA-Seq FASTQ file: 3: ctrl R2 - rev Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33 Mean Inner Distance between Mate Pairs: 150 TopHat settings to use: Default settings Use the Full parameter list to change default settings. Execute

**Tophat Overview**  
Tophat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Please cite: Trapnell, C., Pachter, L., and Salzberg, S.L. Transcript assembly in RNA-seq: Quantitative comparison of methods. Bioinformatics, 2009, 25(33), 1124-1132.

History

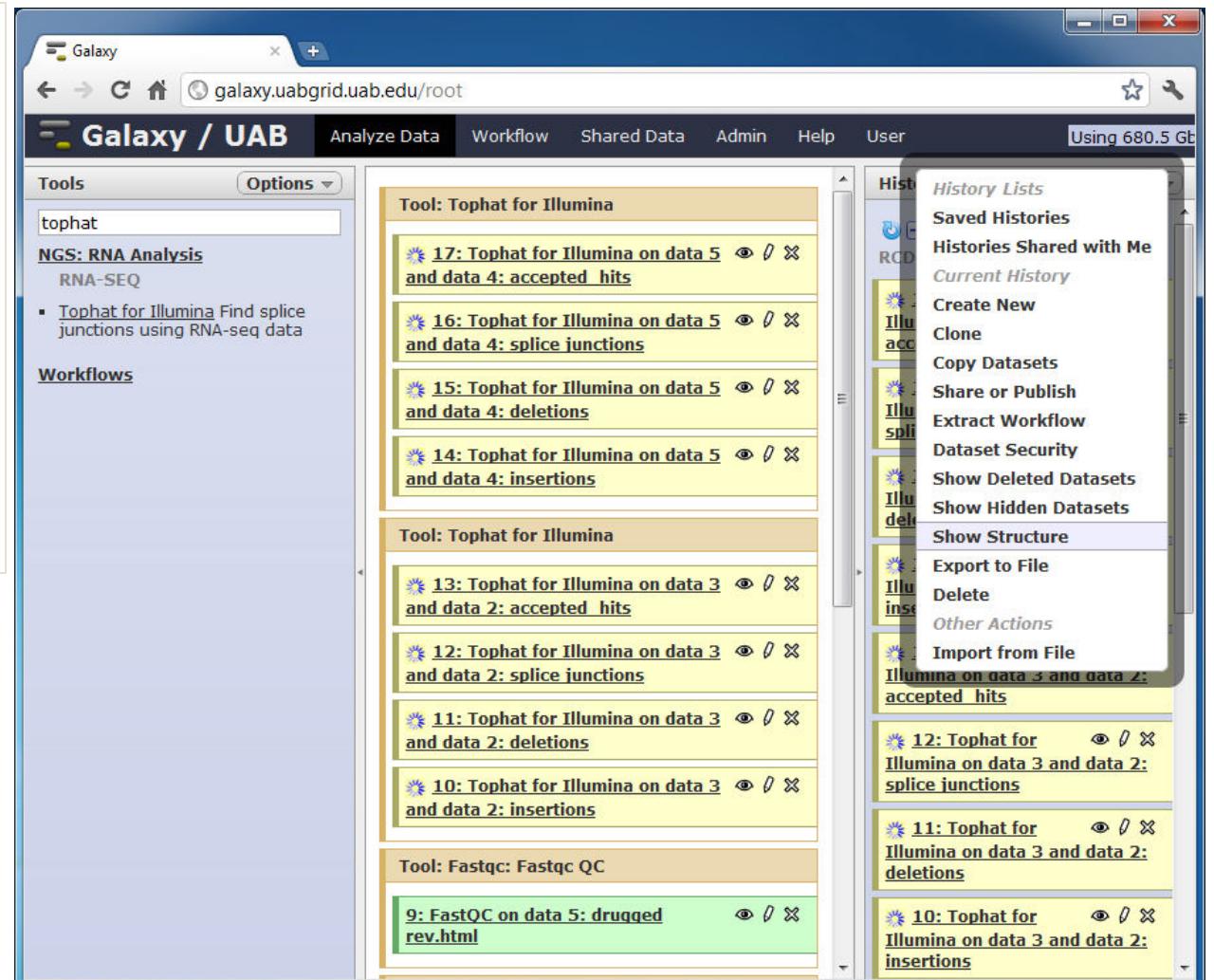
- WS9 RNA-seq Mouse 94.1 Mb
- 10: Tophat for Illumina on data 3 and data 2: accepted hits
- 9: Tophat for Illumina on data 3 and data 2: splice junctions
- 8: Tophat for Illumina on data 3 and data 2: deletions
- 7: Tophat for Illumina on data 3 and data 2: insertions
- 6: FastQCdrugR2\_data 5.html
- 5: drug R2 - rev
- 4: drug R1 - fwd
- 3: ctrl R2 - rev
- 2: ctrl R1 - fwd
- 1: mm9-genes.gtf

# Launched 2 tophats (4 outputs each)

TIP:

 > Show Structure  
groups datasets  
produced by the same  
tool.

 Edit “accepted\_hits”  
Datasets to give shorter  
names with sample info:  
*[ctrl] tophat.accepted\_hits*



# Launch 2 flagstats: QC mapping

TIP:

You can execute a tool before its predecessors have finished running.

Galaxy will queue it until all the dependencies are satisfied!

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

- Tools Panel:** Shows the "flagstat" tool under the "NGS: SAM Tools" category. A blue circle highlights the tool name.
- BAM File to Convert:** A dropdown menu shows three options: "10: [ctrl] tophat.accepted\_hits", "10: [ctrl] tophat.accepted\_hits", and "14: [drug] tophat.accepted\_hits". The third option is highlighted with a blue box.
- Execute Button:** A blue button labeled "Execute" is visible.
- What it does:** A description states: "This tool uses the [SAMTools](#) toolkit to produce simple stats on a BAM file."
- Citation:** A citation for Li et al. (2009) is provided.
- Output:** The results of the executed tool (14) are displayed:

```
76275 in total
0 QC failure
0 duplicates
76275 mapped (100.00%)
76275 paired in sequencing
37766 read1
38509 read2
62200 properly paired (81.55%)
74302 with itself and mate mapped
1973 singlettons (2.59%)
0 with mate mapped to a different chr
0 with mate mapped to a different chr (mapQ>=5)
```

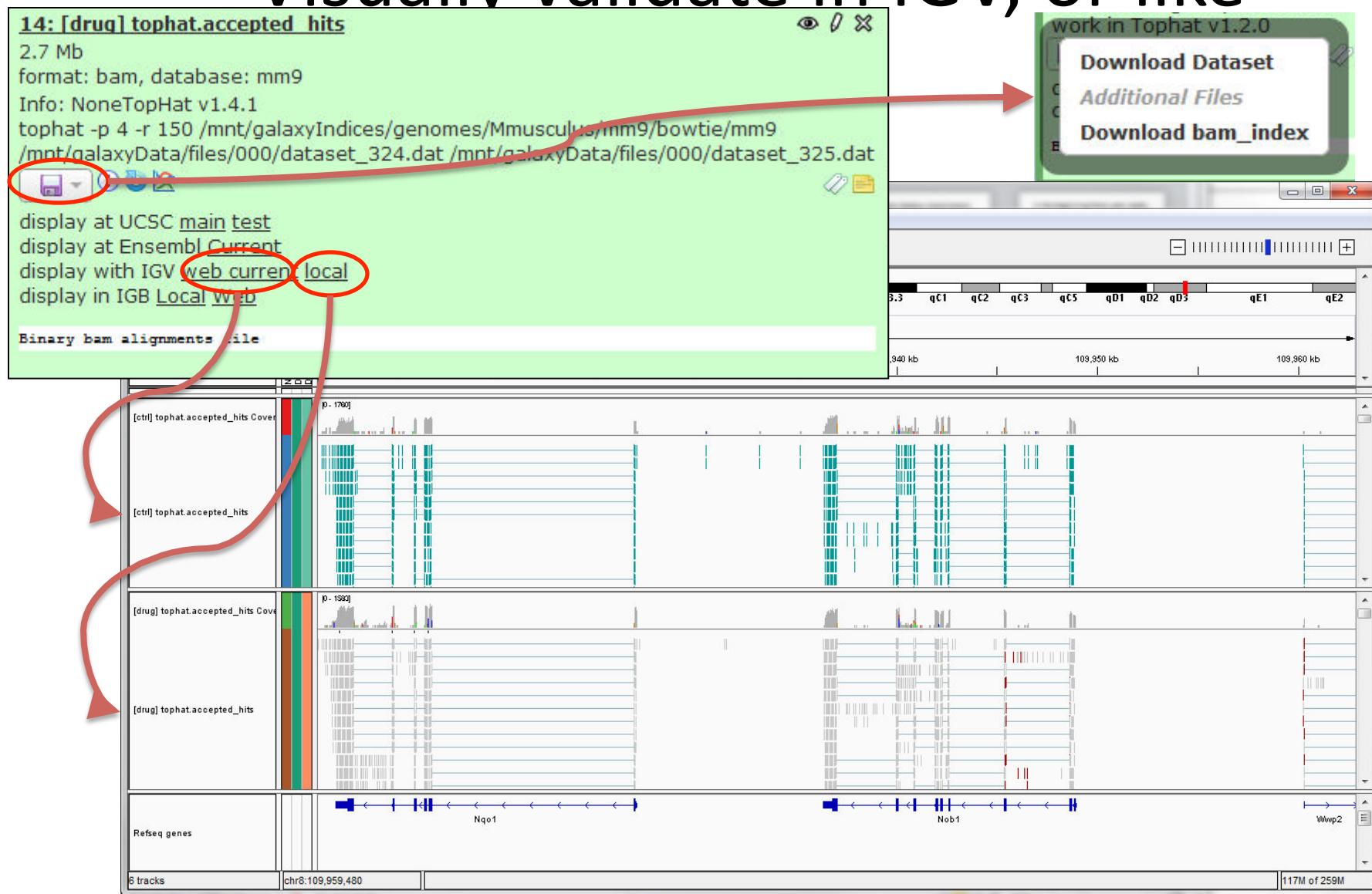
A red circle highlights the first few lines of the output, and a green circle highlights the last few lines. A red arrow points from the green circle to the text "Ignore! Always 100%".
- History:** The history panel shows several previous runs, with the most recent ones being "16: flagstat on data 14" and "15: flagstat on data 10". The "16" run is circled in red.

# Viewing Alignments

- UCSC
- IGV via “web current” (JNLP)
  - Set Visibility window & restart
  - Un-collapse gene track!
- IGV via “local”
  - Make sure it’s running on your machine first!
- Trackster
  - Build visualizations inside galaxy!

# Display mapped BAM & BAI

## Visually validate in IGV, or like



# Cufflink: Construct Transcripts

Read data:  
accepted\_hits (BAM)

**Use Reference Anno = yes**  
then pull down appears you  
can choose a GTF file from  
your history.

Reference annotation:  
iGenome genes.GTF

**Other Parameters**  
*“Perform quartile  
normalization” &  
“Perform Bias Correction”*  
Normally YES is best

because we're not working  
with a full read-set, **use NO** or  
statistics will go haywire

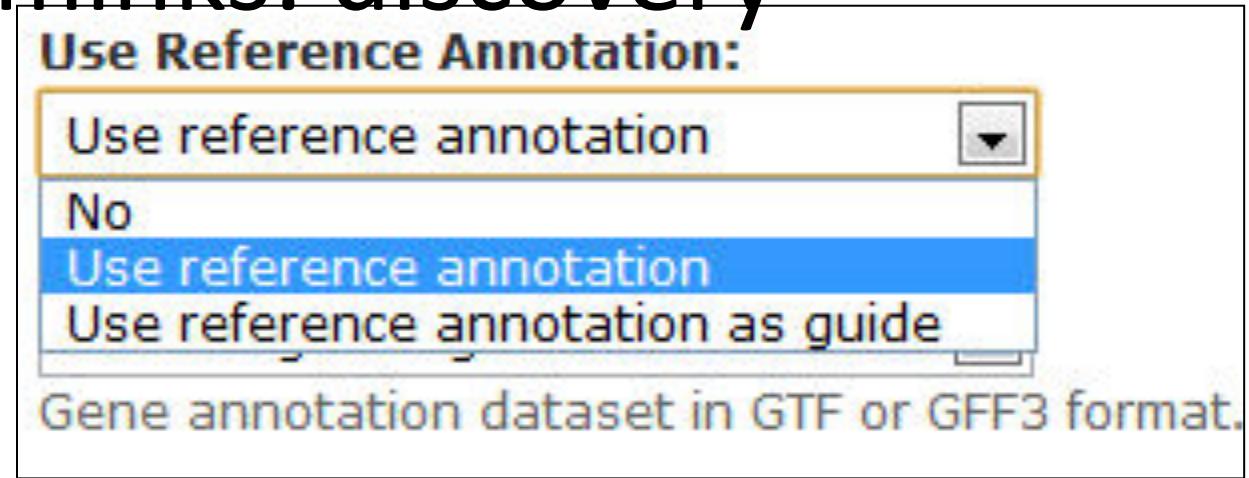
The screenshot shows the Galaxy web interface with the Cufflinks tool selected. The 'Tools' search bar contains 'cufflinks'. The main panel displays the Cufflinks configuration form. Key settings include:

- SAM or BAM file of aligned RNA-Seq reads:** A dropdown menu showing options: 14: [drug] tophat.accepted\_hits, 10: [ctrl] tophat.accepted\_hits, and 14: [drug] tophat.accepted\_hits. The value '300000' is entered below.
- Min Isoform Fraction:** Set to 0.1.
- Pre mRNA Fraction:** Set to 0.15.
- Perform quartile normalization:** Set to **No**. A note explains: "Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts."
- Use Reference Annotation:** Set to **Use reference annotation**.
- Reference Annotation:** A dropdown menu showing option 1: mm9-genes.gtf. A note says: "Gene annotation dataset in GTF or GFF3 format."
- Perform Bias Correction:** Set to **No**. A note says: "Bias detection and correction can significantly improve accuracy of transcript abundance estimates."
- Set Parameters for Paired-end Reads? (not recommended):** Set to **No**.

At the bottom is a blue 'Execute' button. Below it is a 'Cufflinks Overview' section with a progress bar and some text.

The right side of the interface shows the 'History' panel with a list of previous steps, including Tophat runs and flagstat reports for data 14 and 10.

# Cufflinks: discovery



## No – for novel transcript discovery

- requires very deep sequencing – consider sample enrichment!
- Slow, memory intensive
- Discover ONLY transcripts proven by reads

## Ref

- use reads **only** to quantify reference annotation
- All novel splices ignored

## Ref as Guide – most common

- use reference annotation
- Extend reference, when there is sufficient evidence

# Cufflinks outputs

TIP: hide the cmd bar!

assembled\_transcripts (GTF)

list of isoforms

transcript\_expression (tab)

isoforms w/ FPKM

gene\_expression (tab)

genes w/ FPKM

## QC steps

1. Check FPKM not all 0.0!
2. Check you have gene symbols
3. Visualize assembled\_transcripts in a genome browser (UCSC, IGV, IGB, etc)

Galaxy / UAB

Analyze Data Workflow Shared Data Admin Help User Using 680.

History Options

25: Cufflinks on data 17 and data 1: assembled transcripts  
133 lines  
format: gtf, database: mm9  
Info: cufflinks v1.0.3  
cufflinks -q --no-update-check -I 300000 -F 0.050000 -j 0.050000 -p 4  
display at UCSC main test  
display at Ensembl Current

24: Cufflinks on data 17 and data 1: transcript expression

23: Cufflinks on data 17 and data 1: gene expression

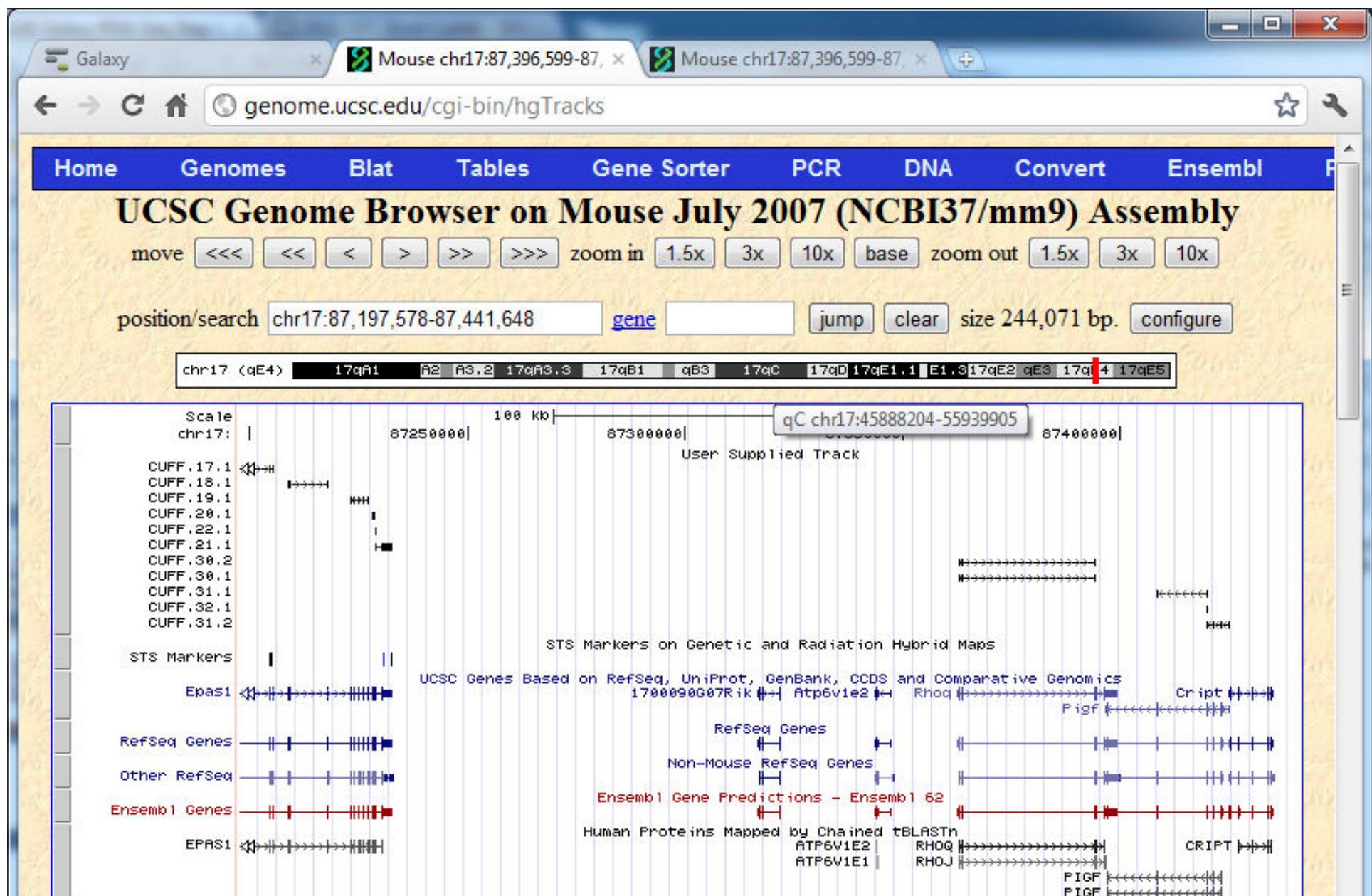
22: Cufflinks on data 13 and data 1: assembled transcripts

Mouse chr17:87,396,599-87

galaxy.uabgrid.uab.edu/root

tracking_id	class_code	nearest_ref_id	gene_id	status	FPKM
FPKM_conf_lo	FPKM_conf_hi				
CUFF.1	-	CUFF.1	-	-	chr17:85176281-
85189887	-	-	OK	29921.5	22130.6 37712.4
CUFF.2	-	CUFF.2	-	-	chr17:85357348-
85393064	-	-	OK	23853.5	17794.7 29912.3
CUFF.3	-	CUFF.3	-	-	chr17:85487793-
85488975	-	-	OK	2206.62	1319.19 3094.05
CUFF.4	-	CUFF.4	-	-	chr17:85489263-
85489585	-	-	OK	19838.5	12236 27440.9
CUFF.5	-	CUFF.5	-	-	chr17:85487793-
85489585	-	-	OK	11468.6	8123.97 14813.3
CUFF.6	-	CUFF.6	-	-	chr17:85489976-
85495959	-	-	OK	5688.73	2090.86 9286.61
CUFF.7	-	CUFF.7	-	-	chr17:85897965-
85898160	-	-	OK	425175	348796 501554
CUFF.8	-	CUFF.8	-	-	chr17:85898262-
85898386	-	-	OK	2.58082e+06	2.130e+06
3.03094e+06					
CUFF.9	-	CUFF.9	-	-	chr17:86183637-
86084828	-	-	OK	11051.1	9074.25 13028
CUFF.10	-	CUFF.10	-	-	chr17:86086757-
86087262	-	-	OK	8992.19	5767.1 12222.3
CUFF.11	-	CUFF.11	-	-	chr17:86887501-
86887644	-	-	OK	187312	9452.4 278172
CUFF.12	-	CUFF.12	-	-	chr17:86892565-
86895452	-	-	OK	23634.9	16588 30680.8
CUFF.13	-	CUFF.13	-	-	chr17:86958492-
87029546	-	-	OK	7428.93	4952.62 9905.23
CUFF.14	-	CUFF.14	-	-	chr17:87053899-
87057181	-	-	OK	8167.22	7201.18 9125.27
CUFF.15	-	CUFF.15	-	-	chr17:87153170-
87153639	-	-	OK	17193.8	12950.9 22636.6
CUFF.16	-	CUFF.16	-	-	chr17:86384075-
86457105	-	-	OK	2.912.9	18140.2 31685.6
CUFF.17	-	CUFF.17	-	-	chr17:87196457-
87205140	-	-	OK	11090.3	6465.3 15715.2

# assembled\_transcripts Visualization



# Cuffmerge: combine transcripts samples & replicates -> complete transcriptome

Run once, using both samples

## Inputs

[ctrl] assembled\_transcripts

[drug] assembled\_transcripts

Reference\_annotation (GTF)

The screenshot shows the Galaxy web interface for the Cuffmerge tool. On the left, the configuration panel for 'Cuffmerge (version 0.0.5)' is displayed. It includes dropdown menus for 'GTF file produced by Cufflinks' (containing entries for 'ctrl' and 'drug'), 'Additional GTF Input Files' (with a 'Remove Additional GTF Input Files' button), 'Use Reference Annotation' (set to 'Yes'), 'Reference Annotation' (set to 'mm9-genes.gtf'), and 'Use Sequence Data' (set to 'No'). Below these are sections for 'Cuffmerge Overview' and 'Know what you are doing'. The 'Overview' section contains a citation for Cuffmerge from Nature Biotechnology. The 'Know what you are doing' section has a warning icon. On the right, the 'History' panel lists 23 completed jobs. The first two entries are circled in red: '23: [drug] Cufflinks on data 14 and data 1: assembled transcripts' and '19: [ctrl] Cufflinks on data 10 and data 1: assembled transcripts'. The bottom entry, '1: mm9-genes.gtf', is also circled in red.

Job ID	Description	Status
23	[drug] Cufflinks on data 14 and data 1: assembled transcripts	✓
22	Cufflinks on data 14 and data 1: transcript expression	✓
21	Cufflinks on data 14 and data 1: gene expression	✓
19	[ctrl] Cufflinks on data 10 and data 1: assembled transcripts	✓
18	Cufflinks on data 10 and data 1: transcript expression	✓
17	Cufflinks on data 10 and data 1: gene expression	✓
16	flagstat on data 14	✓
15	flagstat on data 10	✓
14	[drug] tophat.accepted_hits	✓
13	Tophat for Illumina on data 5 and data 4: splice junctions	✓
12	Tophat for Illumina on data 5 and data 4: deletions	✓
11	Tophat for Illumina on data 5 and data 4: insertions	✓
10	[ctrl] tophat.accepted_hits	✓
9	Tophat for Illumina on data 3 and data 2: splice junctions	✓
8	Tophat for Illumina on data 3 and data 2: deletions	✓
7	Tophat for Illumina on data 3 and data 2: insertions	✓
6	FastQCdrugR2_data 5.html	✓
5	drug R2 - rev	✓
4	drug R1 - fwd	✓
3	ctrl R2 - rev	✓
2	ctrl R1 - fwd	✓
1	mm9-genes.gtf	✓

# Cuffdiff: Quantize Transcripts

## Compute fold change between conditions

Run once,  
use all samples & replicates

ALWAYS create replicate  
groups for each condition

ALWAYS name your groups

even if you have only one  
replicate!

The screenshot shows the Galaxy web interface for running Cuffdiff. On the left, the 'Cuffdiff (version 0.0.5)' tool configuration is displayed. It includes fields for 'Transcripts' (set to '32: Cuffmerge on data..transcript'), 'Perform replicate analysis' (set to 'Yes'), and two 'Groups' sections: 'Group 1' (name 'ctrl') and 'Group 2' (name 'drug'). Each group has a 'Replicates' section with a dropdown menu showing '10: [ctrl] tophat.accepted\_hits' and '14: [drug] tophat.accepted\_hits'. On the right, the 'History' panel lists various workflow steps, with several entries circled in red: '32: Cuffmerge on data 19, data 1, and data 23: merged transcripts', '14: [drug] tophat.accepted\_hits' (circled twice), '10: [ctrl] tophat.accepted\_hits' (circled twice), and '3: ctrl R2 - rev'.

Step ID	Description	Status
32	Cuffmerge on data..transcript	Completed
23	[drug] Cufflinks on data 14 and data 1: assembled transcripts	Completed
22	Cufflinks on data 14 and data 1: transcript expression	Completed
21	Cufflinks on data 14 and data 1: gene expression	Completed
19	[ctrl] Cufflinks on data 10 and data 1: assembled transcripts	Completed
18	Cufflinks on data 10 and data 1: transcript expression	Completed
17	Cufflinks on data 10 and data 1: gene expression	Completed
16	flagstat on data 14	Completed
15	flagstat on data 10	Completed
14	[drug] tophat.accepted_hits	Completed
13	Tophat for Illumina on data 5 and data 4: splice junctions	Completed
12	Tophat for Illumina on data 5 and data 4: deletions	Completed
11	Tophat for Illumina on data 5 and data 4: insertions	Completed
10	[ctrl] tophat.accepted_hits	Completed
9	Tophat for Illumina on data 3 and data 2: splice junctions	Completed
8	Tophat for Illumina on data 3 and data 2: deletions	Completed
7	Tophat for Illumina on data 3 and data 2: insertions	Completed
6	fastQCdrugR2_data 5.html	Completed
5	drug R2 - rev	Completed
4	drug R1 - fwd	Completed
3	ctrl R2 - rev	Completed

# Cuffdiff: Fold Change Between Conditions

Run once, using both samples

ALWAYS create replicate groups for each condition,  
even if you have only one replicate!

Short names – no spaces or special characters!!!

Galaxy / UAB

Groups

Group 1

Group name (no spaces or commas): control

Replicates

Replicate 1

Add file: 13: Tophat for Illumi..ts: control

Remove Replicate 1

Add new Replicate

Remove Group 1

Group 2

Group name (no spaces or commas): drugged

Replicates

Replicate 1

Add file: 17: Tophat for Illumi..ts: drugged

Remove Replicate 1

Add new Replicate

Remove Group 2

History

- 22: Cufflinks on data 13 and data 1: assembled transcripts
- 21: Cufflinks on data 13 and data 1: transcript expression
- 20: Cufflinks on data 13 and data 1: gene expression
- 19: flagstat on data 17
- 18: flagstat on data 13: control
- 17: Tophat for Illumina on data 5 and data 4: accepted hits: drugged
- 16: Tophat for Illumina on data 5 and data 4: splice junctions
- 15: Tophat for Illumina on data 5 and data 4: deletions
- 14: Tophat for Illumina on data 5 and data 4: insertions
- 13: Tophat for Illumina on data 3 and data 2: accepted hits: control
- 12: Tophat for Illumina on data 3 and data 2: splice junctions
- 11: Tophat for Illumina on data 3 and data 2: deletions
- 10: Tophat for Illumina on data 3 and data 2: insertions
- 9: FastQC on data 5: drugged rev.html
- 8: FastQC on data 4: drugged fwd.html
- 7: FastQC on data 3: control rev.html
- 6: FastQC on data 2: control fwd.html
- 5: drugged\_mm9\_chr15\_Plekhh2-PigF\_reverse.fastq

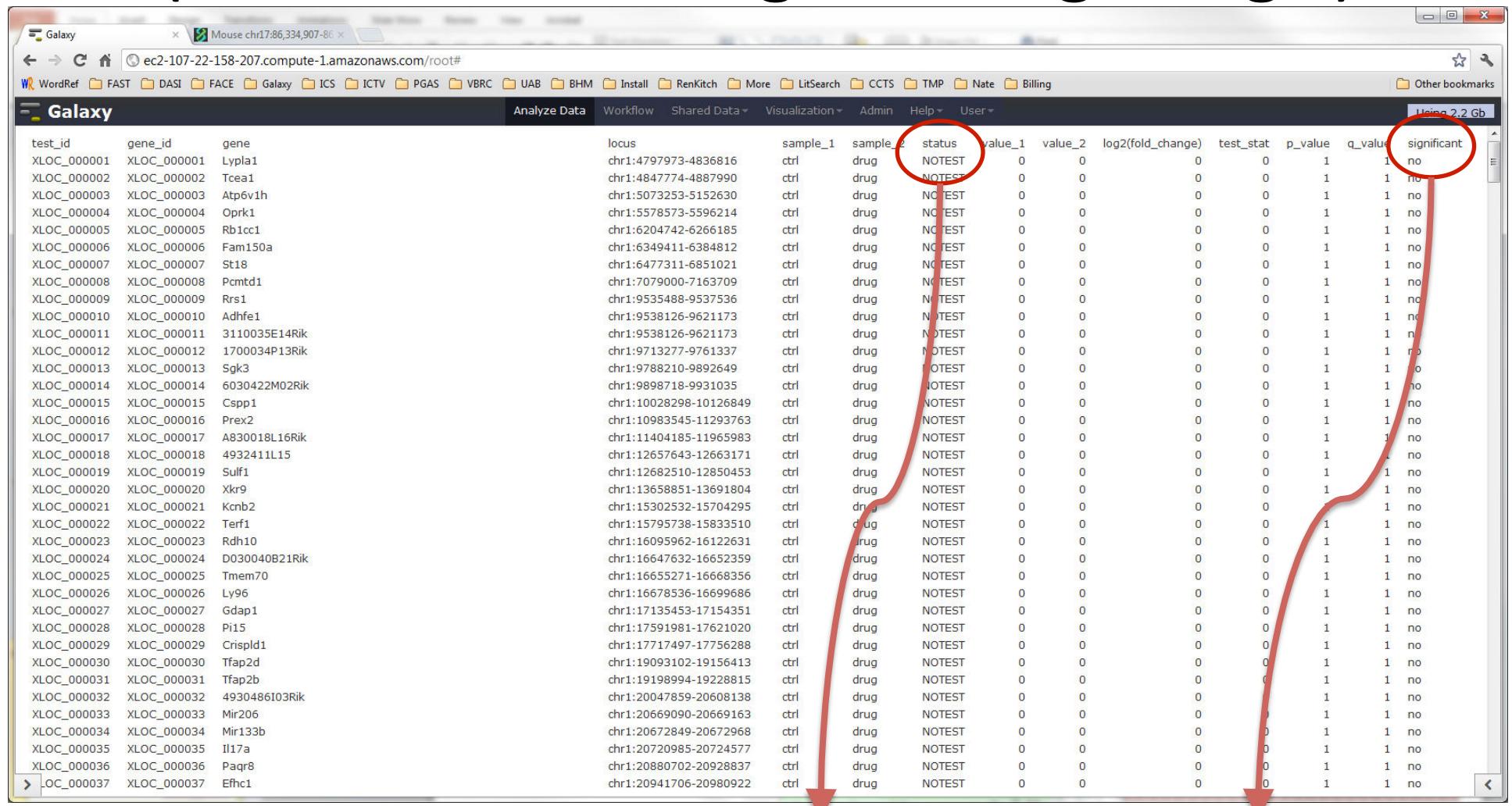
# Cuffdiff output ( > view structure)

Tool: Cuffdiff		
<a href="#">43: Cuffdiff on data 14, data 10, and data 32: transcript FPKM tracking</a>	①	✗
<a href="#">42: Cuffdiff on data 14, data 10, and data 32: transcript differential expression testing</a>	①	✗
<a href="#">41: Cuffdiff on data 14, data 10, and data 32: gene FPKM tracking</a>	①	✗
<a href="#">40: Cuffdiff on data 14, data 10, and data 32: gene differential expression testing</a>	①	✗
<a href="#">39: Cuffdiff on data 14, data 10, and data 32: TSS groups FPKM tracking</a>	①	✗
<a href="#">38: Cuffdiff on data 14, data 10, and data 32: TSS groups differential expression testing</a>	①	✗
<a href="#">37: Cuffdiff on data 14, data 10, and data 32: CDS FPKM tracking</a>	①	✗
<a href="#">36: Cuffdiff on data 14, data 10, and data 32: CDS FPKM differential expression testing</a>	①	✗
<a href="#">35: Cuffdiff on data 14, data 10, and data 32: CDS overloading differential expression testing</a>	①	✗
<a href="#">34: Cuffdiff on data 14, data 10, and data 32: promoters differential expression testing</a>	①	✗
<a href="#">33: Cuffdiff on data 14, data 10, and data 32: splicing differential expression testing</a>	①	✗
Tool: Cuffmerge		
<a href="#">32: Cuffmerge on data 19, data 1, and data 23: merged transcripts</a>	①	✗
Tool: Cufflinks		

FILTER!!

# Cuffdiff “gene diff”: 23,113 lines!

(  -> value for all genes in gene.gtf)



test_id	gene_id	gene
XLOC_000001	XLOC_000001	Lypla1
XLOC_000002	XLOC_000002	Tcea1
XLOC_000003	XLOC_000003	Atp6v1h
XLOC_000004	XLOC_000004	Oprk1
XLOC_000005	XLOC_000005	Rb1cc1
XLOC_000006	XLOC_000006	Fam150a
XLOC_000007	XLOC_000007	St18
XLOC_000008	XLOC_000008	Pcmtd1
XLOC_000009	XLOC_000009	Rrs1
XLOC_000010	XLOC_000010	Adhfe1
XLOC_000011	XLOC_000011	3110035E14Rik
XLOC_000012	XLOC_000012	1700034P13Rik
XLOC_000013	XLOC_000013	Sgk3
XLOC_000014	XLOC_000014	6030422M02Rik
XLOC_000015	XLOC_000015	Cspp1
XLOC_000016	XLOC_000016	Prex2
XLOC_000017	XLOC_000017	A830018L16Rik
XLOC_000018	XLOC_000018	493241L15
XLOC_000019	XLOC_000019	Sulf1
XLOC_000020	XLOC_000020	Xkr9
XLOC_000021	XLOC_000021	Kcnb2
XLOC_000022	XLOC_000022	Terf1
XLOC_000023	XLOC_000023	Rdh10
XLOC_000024	XLOC_000024	D030040B21Rik
XLOC_000025	XLOC_000025	Tmem70
XLOC_000026	XLOC_000026	Ly96
XLOC_000027	XLOC_000027	Gdap1
XLOC_000028	XLOC_000028	Pi15
XLOC_000029	XLOC_000029	Crispld1
XLOC_000030	XLOC_000030	Tfap2d
XLOC_000031	XLOC_000031	Tfap2b
XLOC_000032	XLOC_000032	4930486I03Rik
XLOC_000033	XLOC_000033	Mir206
XLOC_000034	XLOC_000034	Mir133b
XLOC_000035	XLOC_000035	Ii17a
XLOC_000036	XLOC_000036	Paqrl8
XLOC_000037	XLOC_000037	Efhc1

locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
chr1:4797973-4836816	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:4847774-4887990	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:5073253-5152630	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:5578573-5596214	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:6204742-6266185	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:6349411-6384812	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:6477311-6851021	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:7079000-7163709	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:9535488-9537536	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:9538126-9621173	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:9538126-9621173	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:9713277-9761337	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:9788210-9892649	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:9898718-9931035	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:10028298-10126849	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:10983545-11293763	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:11404185-11965983	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:12657643-12663171	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:12682510-12850453	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:13658851-13691804	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:15302532-15704295	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:15795738-15833510	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:16095962-16122631	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:16647632-16652359	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:16655271-16668356	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:16678536-16699686	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:17135453-17154351	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:17591981-17621020	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:17717497-17756288	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:19093102-19156413	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:19198994-19228815	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:20047859-20608138	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:20669090-20669163	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:20672849-20672968	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:20720985-20724577	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:20880702-20928837	ctrl	drug	NOTEST	0	0	0	0	1	1	no
chr1:20941706-20980922	ctrl	drug	NOTEST	0	0	0	0	1	1	no

Tools > Filter and Sort > Filter: c7 <> 'NOTEST' and c14 <> 'no'

# Gene diff: genes with enough reads

## 14 genes

Screenshot of the Galaxy web interface showing the results of a gene filtering process.

The browser title is "Galaxy" and the tab is "Mouse chr17:86,334,907-86". The URL is "ec2-107-22-158-207.compute-1.amazonaws.com/root#".

The main menu includes: Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, User.

The "Using 2.2 Gb" message is displayed in the top right.

**Tools** sidebar:

- Get Data
- Send Data
- ENCODE Tools
- Lift-Over
- Text Manipulation
- Filter and Sort**
  - Filter data on any column using simple expressions
  - Sort data in ascending or descending order
  - Select lines that match an expression
- GFF
- Extract features from GFF data
- Filter GFF data by attribute using simple expressions
- Filter GFF data by feature count using simple expressions
- Filter GTF data by attribute values list

**Filter (version 1.1.0)** panel:

**Filter:** 43: Cuffdiff on data ..KM tracking  
Dataset missing? See TIP below.

**With following condition:** `c7 <> 'NOTESt'` (highlighted with a red arrow)

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

**Execute** button.

**TIP:** Double equal signs, ==, must be used as "equal to" (e.g., `c1 == 'chr22'`)

**TIP:** Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

**TIP:** If your data is not TAB delimited, use Text Manipulation->Convert

**Syntax**: The filter tool allows you to restrict the dataset using simple conditional statements. Columns are referenced with c and a number.

**History** panel:

- WS9 RNA-seq Mouse (379.7 Mb)
- 44: Filter on data 40
- 43: Cuffdiff on data 14, data 10, and data 32: transcript FPKM tracking
- 42: Cuffdiff: transcript diff
- 41: Cuffdiff on data 14, data 10, and data 32: gene FPKM tracking
- 40: Cuffdiff: gene diff**
  - 23,113 lines
  - format: tabular, database: mm9
  - Info: cuffdiff v1.3.0 (3022)  
cuffdiff --no-update-check -q -p 4 -c 10 --FDR 0.050000 -l labels ctrl drug /mnt/galaxyData/files/000/dataset\_414.dat /mnt/galaxyData/files/000/dataset\_382.dat /mnt/galaxyData/files/000/dataset\_396.dat
- 39: Cuffdiff on data 14, data 10, and data 32: TSS groups FPKM tracking
- 38: Cuffdiff on data 14, data 10, and data 32: TSS groups differential expression testing
- 37: Cuffdiff on data 14, data 10, and data 32: CDS FPKM tracking

A red circle highlights the value "NOTESt" in the 7th column of the 40: Cuffdiff: gene diff table row for XLOC\_000001.

1	2	3	4	5	6	7	8	9	10	11	12
test_id	gene_id	gene	locus		sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat p_val
XLOC_000001	XLOC_000001	Lyp1a1	chr1:4797973-4836816	ctrl	drug	NOTESt	0	0	0	0	1
XLOC_000002	XLOC_000002	Tceal1	chr1:4847774-4887990	ctrl	drug	NOTESt	0	0	0	0	1
XLOC_000003	XLOC_000003	Atp6vlh	chr1:5073253-5152630	ctrl	drug	NOTESt	0	0	0	0	1
XLOC_000004	XLOC_000004	Oprk1	chr1:5578573-5596214	ctrl	drug	NOTESt	0	0	0	0	1
XLOC_000005	XLOC_000005	Rbiccl	chr1:6204742-6266185	ctrl	drug	NOTESt	0	0	0	0	1

# Gene diff: statistically significant 9 genes

**Galaxy**      Analyze Data      Workflow      Shared Data      Visualization      Admin      Help      User      Using 2.2 Gb

**Filter (version 1.1.0)**

**Filter:**  
44: cuffdiff.genes <> 'NOTEST'  
Dataset missing? See TIP below.

**With following condition:**  
c14 <> 'no'

Double equal signs, ==, must be used as shown above. To filter for an arbitrary string, use the Select tool.

**Execute**      2

**TIP:** Double equal signs, ==, must be used as "equal to" (e.g., c1 == 'chr22')

**TIP:** Attempting to apply a filtering condition may throw exceptions if the data type (e.g., string, integer) in every line of the columns being filtered is not appropriate for the condition (e.g., attempting certain numerical calculations on strings). If an exception is thrown when applying the condition to a line, that line is skipped as invalid for the filter condition. The number of invalid skipped lines is documented in the resulting history item as a "Condition/data issue".

**TIP:** If your data is not TAB delimited, use Text Manipulation->Convert

**Syntax**

The filter tool allows you to restrict the dataset using simple conditional statements. Columns are referenced with **c** and a **number**. For example, **c1** refers to the first column of a tab-delimited file. Make sure that multi-character operators contain no white space ( e.g., <= is valid while < = is not valid ) When using 'equal-to' operator **double equal sign '==' must be used** ( e.g., **c1=='chr1'** ) Non-numerical values must be included in single or double quotes ( e.g., **c6=='+'** ) Filtering condition can include logical operators, but **make sure operators are all lower case** ( e.g., **(c1=='chrX' and c1!= 'chrY') or not c6=='+'** )

**Example**

**c1=='chr1'** selects lines in which the first column is chr1  
**c3-c2<100\*c4** selects lines where subtracting column 3 from column 2 is less than the value of column 4 times 100  
**i(c2.split(',')) < 4** will select lines where the second column has less than four mm9 generated elements

**History**

WS9 RNA-seq Mouse      379.7 Mb

45: cuffdiff.genes significant

9 lines  
format: tabular, database: mm9  
Info: None

1	2	3	4	5	6	7	8	9	10	11
test_id	gen_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat
XLOC_002294	XLOC_002294	Rps26	chr10:128061584-128063562	ctrl	drug	OK	477526	0	-1.79769e+308	-1.79769e+3
XLOC_008261	XLOC_008261	Rps2	chr17:24857007-24858872	ctrl	drug	OK	475.423	29968.6	5.9781	-4.16766
XLOC_008926	XLOC_008926	Lps18	chr17:34088943-34092586	ctrl	drug	OK	642583	16838.4	-5.25406	6.13259
XLOC_017500	XLOC_017500	Gm6554	chr6:146591100-146599122	ctrl	drug	OK	146499	0	-1.79769e+308	-1.79769e+3
XLOC_020075	XLOC_020075	Cyp5b	chr8:109674560-109711370	ctrl	drug	OK	56804.4	153148	1.43085	-2.61382

44: cuffdiff.genes <> 'NOTEST'

14 lines  
format: tabular, database: mm9  
Info: Filtering with c7 <> 'NOTEST', kept 0.06% of 23113 valid lines (23113 total lines).

5	6	7	8	9	10	11	12	13	14	significant
sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	u_value	significant
28061584-128063562	ctrl	drug	477526	0	-1.79769e+308	-1.79769e+308	2.3698e-08	1.54041e-7	yes	
#857007-24858872	ctrl	drug	475.423	29968.6	5.9781	-4.16766	3.07741e-05	0.000100016	yes	
#088943-34092586	ctrl	drug	642583	16838.4	-5.25406	6.13259	8.646e-10	1.12398e-08	yes	
5591100-146599122	ctrl	drug	146499	0	-1.79769e+308	-1.79769e+308	1.74701e-05	7.57038e-05	yes	
5580776-109584828	ctrl	drug	12359	4567.92	-1.49595	1.4195	0.155754	0.20248	no	

43: Cuffdiff on data 14, data 10, and data 32: transcript FPKM tracking

# “Final” gene list

Galaxy / GCC X				Analyze Data		Workflow	Shared Data	Visualization	Admin	Help	User	Using 1.9 Gb		
test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significance	
XLOC_008261	XLOC_008261	Rps2	chr17:24857007-24858872	ctrl	drug	OK	12154.7	20346.7	0.743285	-1.31836	0.187382	0.218612	no	
XLOC_020075	XLOC_020075	Cyb5b	chr8:109674560-109711370	ctrl	drug	OK	73935.3	103978	0.491939	-1.99412	0.0461384	0.0645938	no	
XLOC_020077	XLOC_020077	Wwp2	chr8:109960297-110082495	ctrl	drug	OK	39072.3	3287.69	-3.571	7.57348	3.64153e-14	2.54907e-13	yes	
XLOC_020591	XLOC_020591	Terf2	chr8:109593301-109620489	ctrl	drug	OK	68292.1	106354	0.639088	-3.16016	0.00157683	0.00275946	yes	
XLOC_020592	XLOC_020592	Nqo1	chr8:109912124-109927105	ctrl	drug	OK	0	14465.2	1.79769e+308	1.79769e+308	3.16712e-05	0.000110849	yes	
XLOC_020594	XLOC_020594	Psmd7	chr8:110104279-110112382	ctrl	drug	OK	54160.4	23570.5	-1.20026	3.54574	0.000391515	0.000913536	yes	
XLOC_020696	XLOC_020696	-	chr8:109786639-109787183	ctrl	drug	OK	876105	907362	0.0505752	-0.201707	0.840146	0.840146	no	

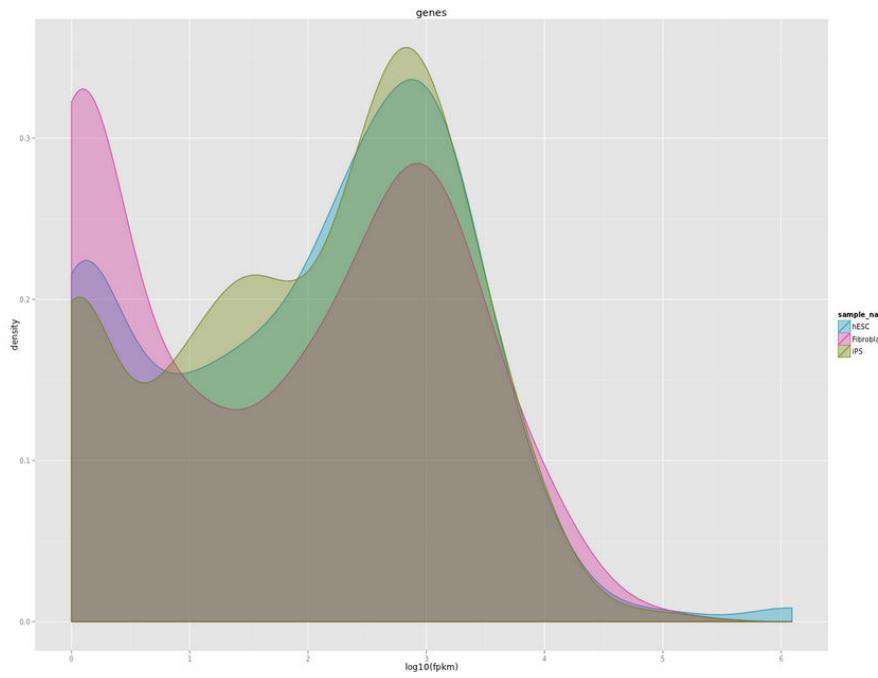
# Download Excel Conditional Formatting

# Visualization of Gene Expression cummeRbund in Galaxy

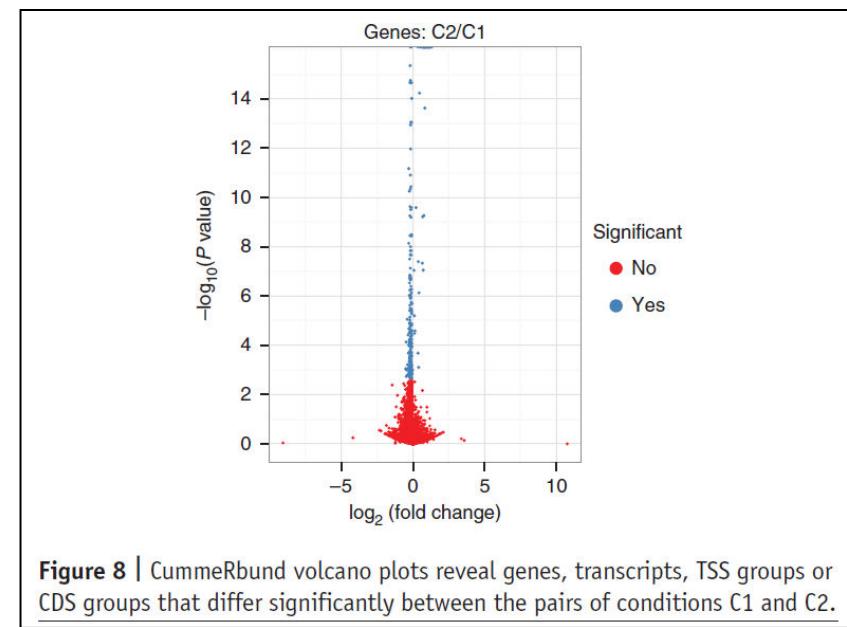
Galaxy wrapper created, but not yet in tool shed

<http://cvrgrid.org/node/235>

<http://ec2-23-20-5-163.compute-1.amazonaws.com:8080/u/boliu/h/cummerbund>



(courtesy of Liu Bo @ University of Chicago)



**Figure 8** | CummeRbund volcano plots reveal genes, transcripts, TSS groups or CDS groups that differ significantly between the pairs of conditions C1 and C2.

(Trapnell et al, 2012, Nat Protocols)

# 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
- Illumina iGenomes: indexes and annotations for use with Cufflinks, etc.
  - <http://cufflinks.cbcb.umd.edu/igenomes.html>

Thanks! Questions? Contact info:

R. Curtis Hendrickson

Research Associate

Microbiology

Center for Clinical and Translational Science

University of Alabama at Birmingham

<http://www.uab.edu/ccts/ResearchResources/BMI>  
curtish arabesque uab.edu