GCC Workshop 9
RNA-Seq with Galaxy

Curtis Hendrickson (curtish@uab.edu)
David Crossman (dkcrossm@uab.edu)
Jeremy Goecks (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

2

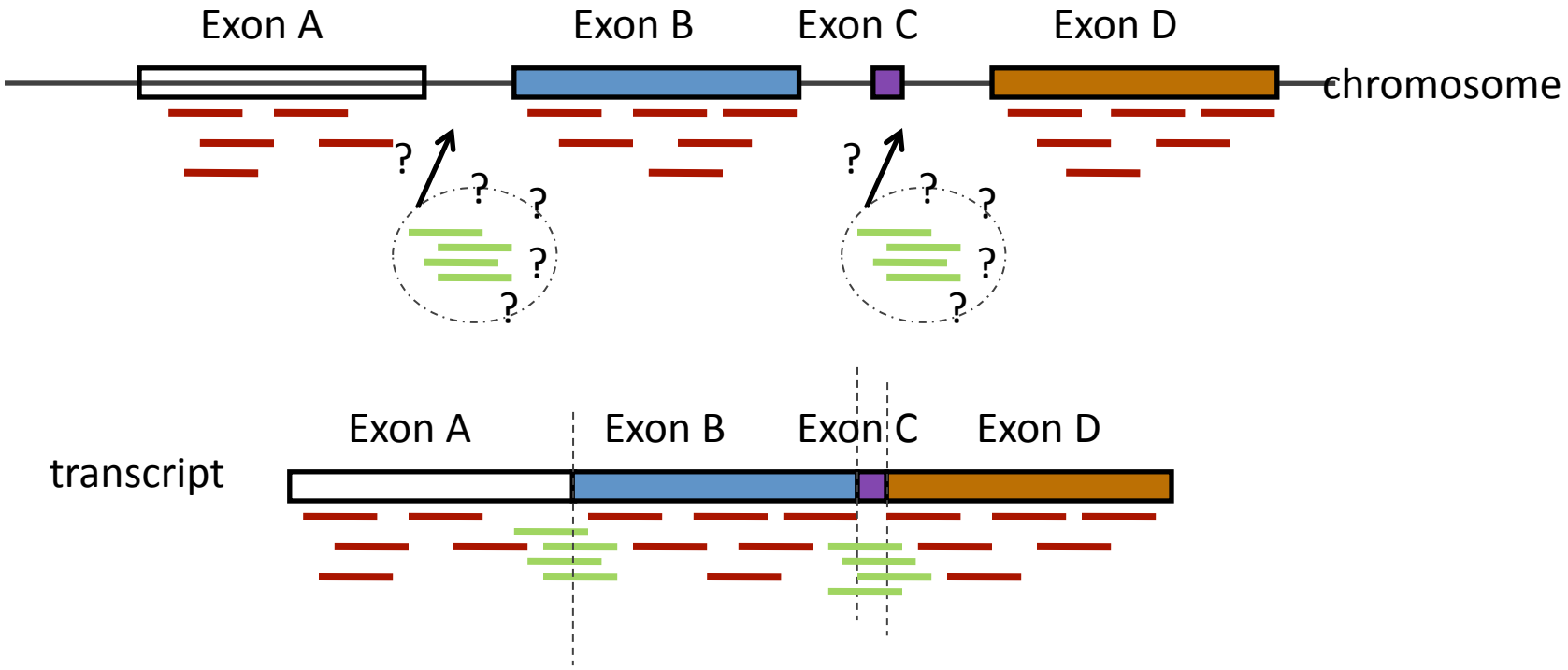
Why RNA-Seq?

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

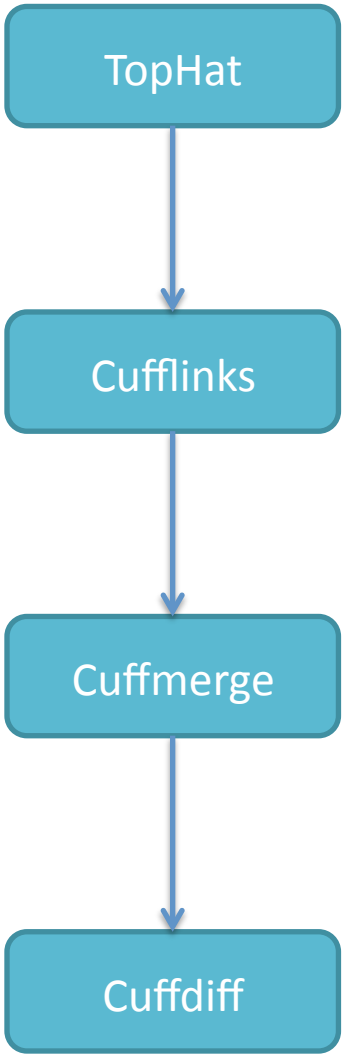
What type of analysis?

- Compare gene or transcript levels between samples or conditions?
  - Sequence the whole transcriptome
- Discovery of alternate splice forms?
  - >= 75bp 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 it's splice junction – not enough to detect.

Reads in RNA-seq



An RNA-Seq Analysis Pipeline



Align reads & discover transcriptome

Assemble reads into isoforms

Build consensus transcriptome across samples and replicates

Quantify and compare transcript expression levels

See *Trapnell et al, 2012, Nat Protocols*

PROTOCOL

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

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

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

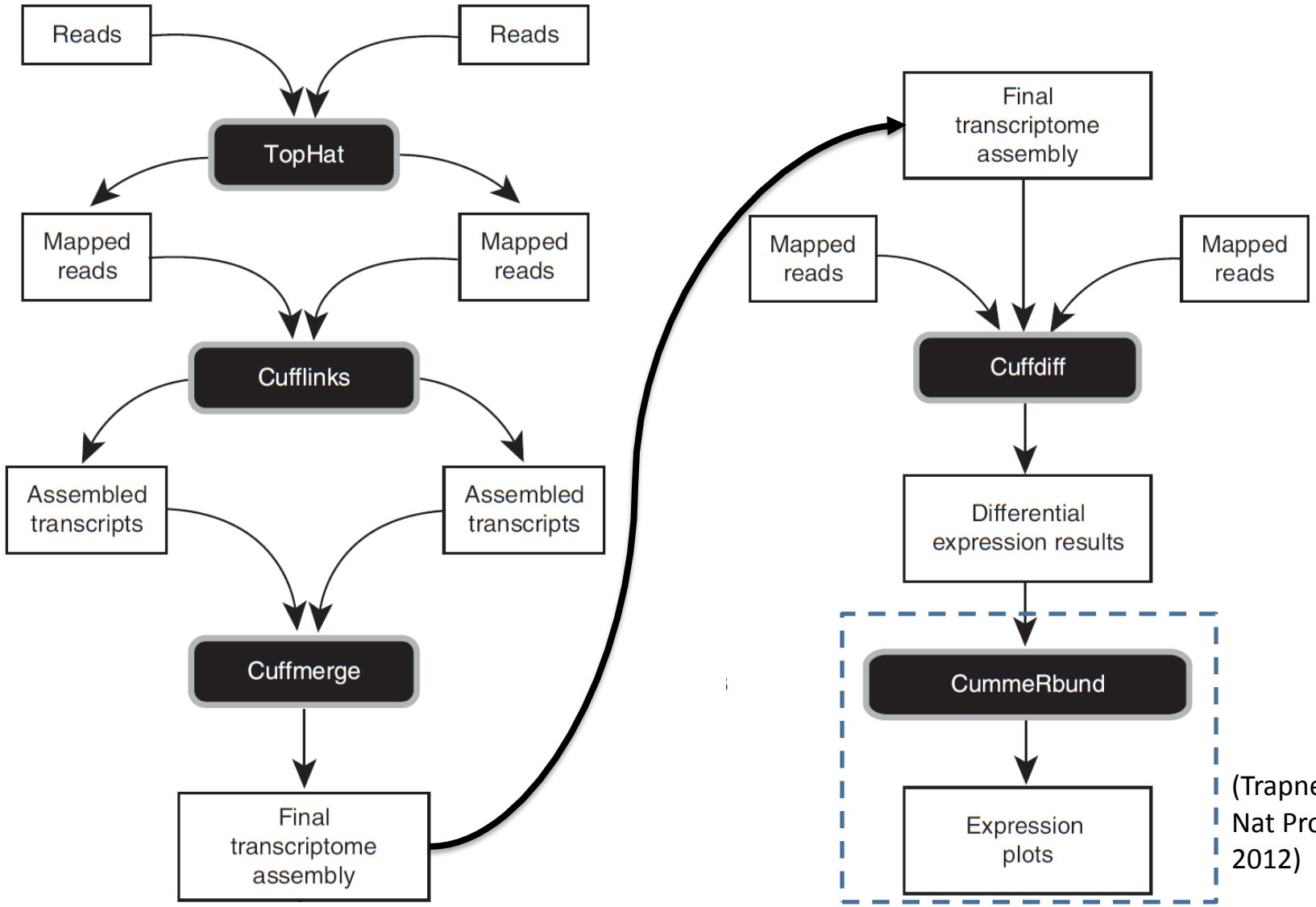
Published online 1 March 2012; doi:10.1038/nprot.2012.016

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

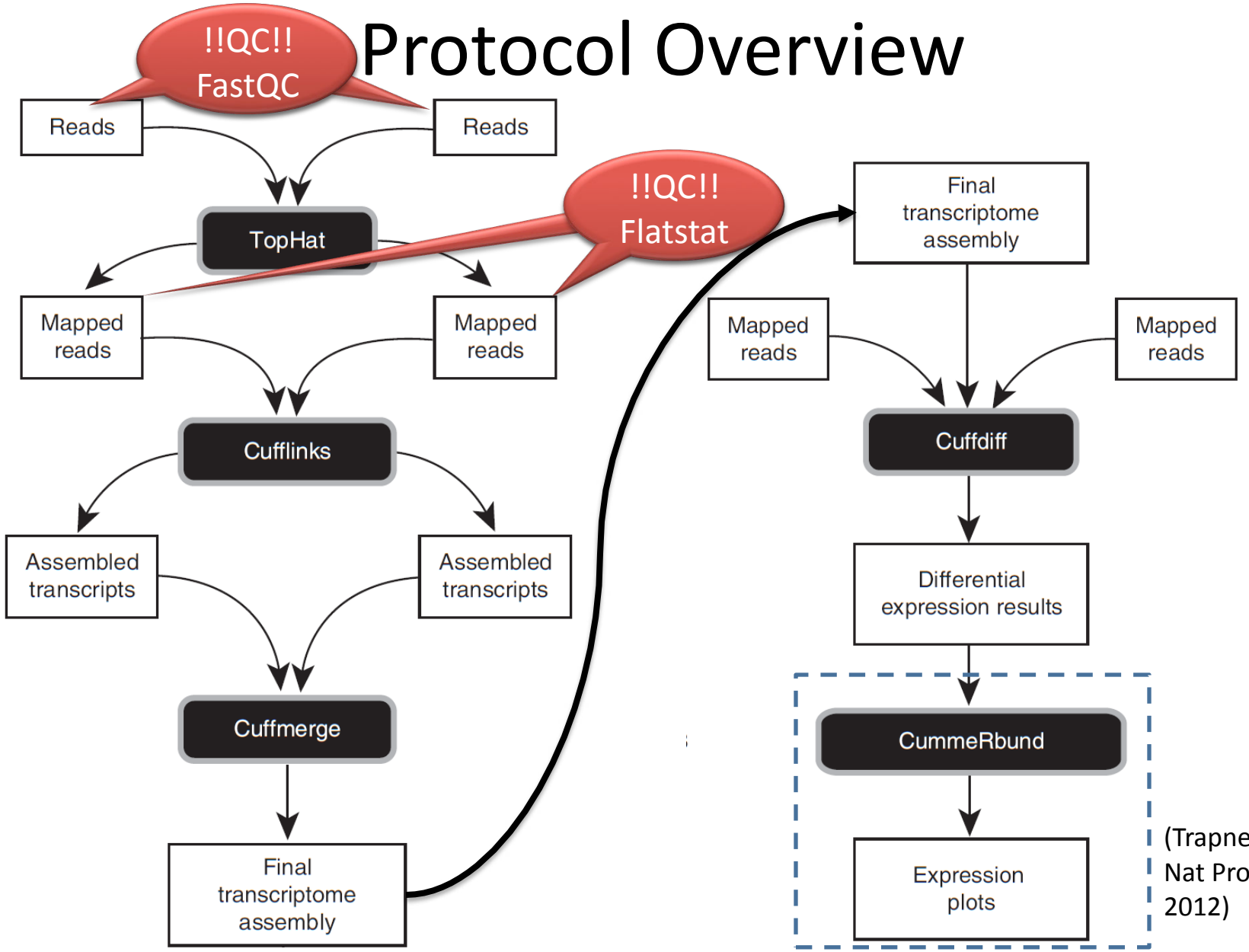
- FASTA <1G Genomes
- FASTQ 1-12G/ea Sequencing reads
- BAM 1-5G/ea Mapped Reads (*binary*)
- BED <1G Genomic intervals
- GTF < 1G Gene/transcript location

2-Condition Differential Expression Protocol Overview



(Trapnell et al, Nat Protocols, 2012)

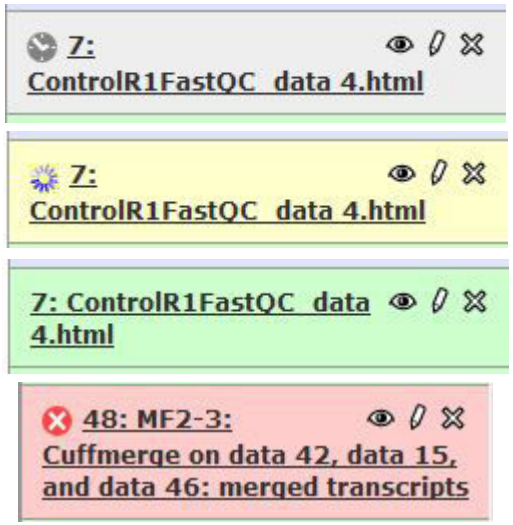
2-Condition Differential Expression Protocol Overview



(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

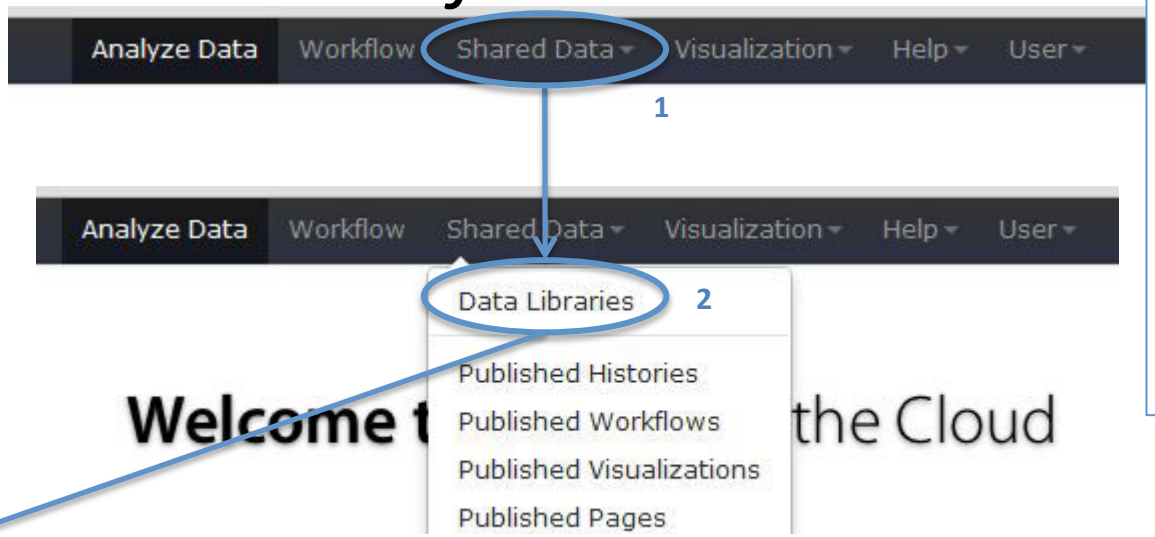
Run this job again

View in Trackster

Edit dataset tags

In the beginning there were reads…

Get them from Shared Data



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)

Data Library “WS9: RNA-Seq Analysis with Galaxy”

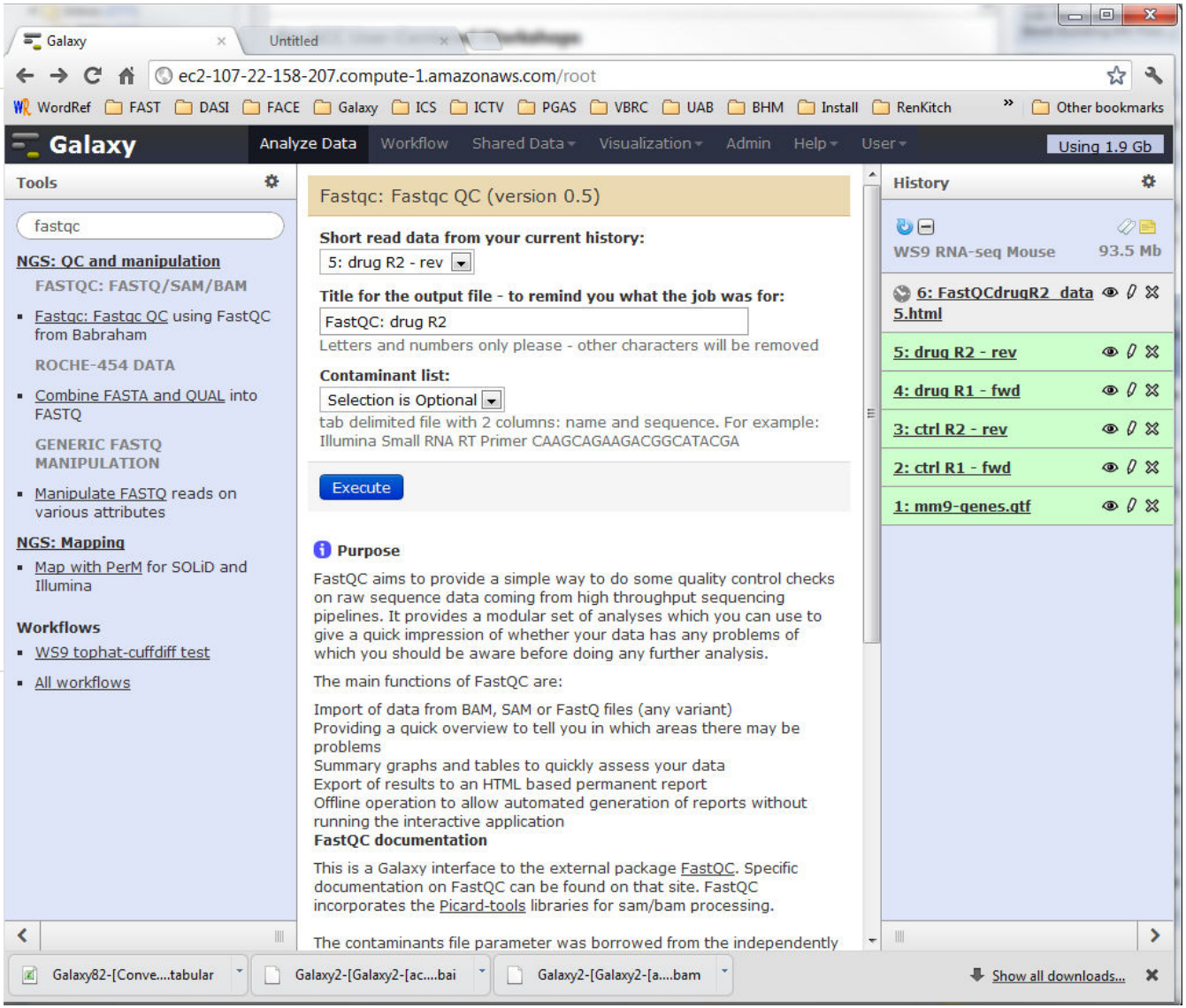
Add datasets | Add folder | Library Actions

Two exercises: mouse and treeshrew

Name	Message	Data type	Date uploaded	File size
mouse	FASTQ and .GTF files for a 2-condition (drug vs ctrl) mouse experiment.			
gene annotation				
mm9-genes.gtf	iGenome/Mus_musculus/UCSC/mm9/genes.gtf See http://cufflinks.cbcb.umd.edu/igenomes.html 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 Keeiling, PhD at University of Alabama at Birmingham (http://www.microbio.uab.edu/faculty/keeling/index.html)			
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

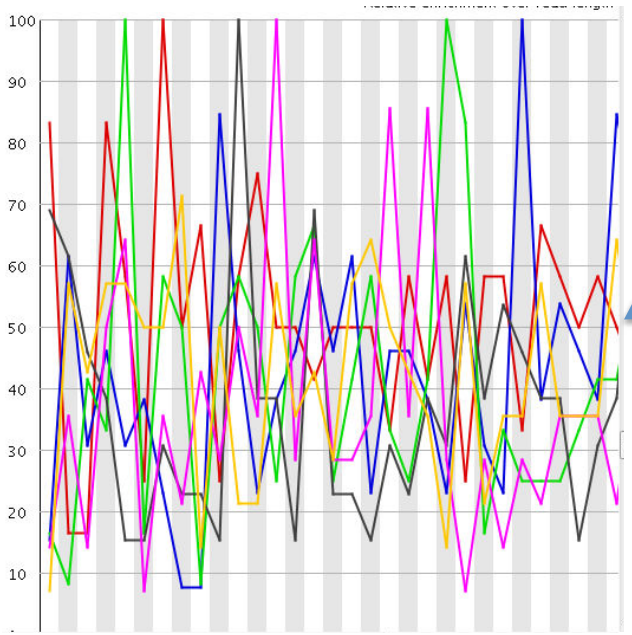
So we used FASTQC to QC them…

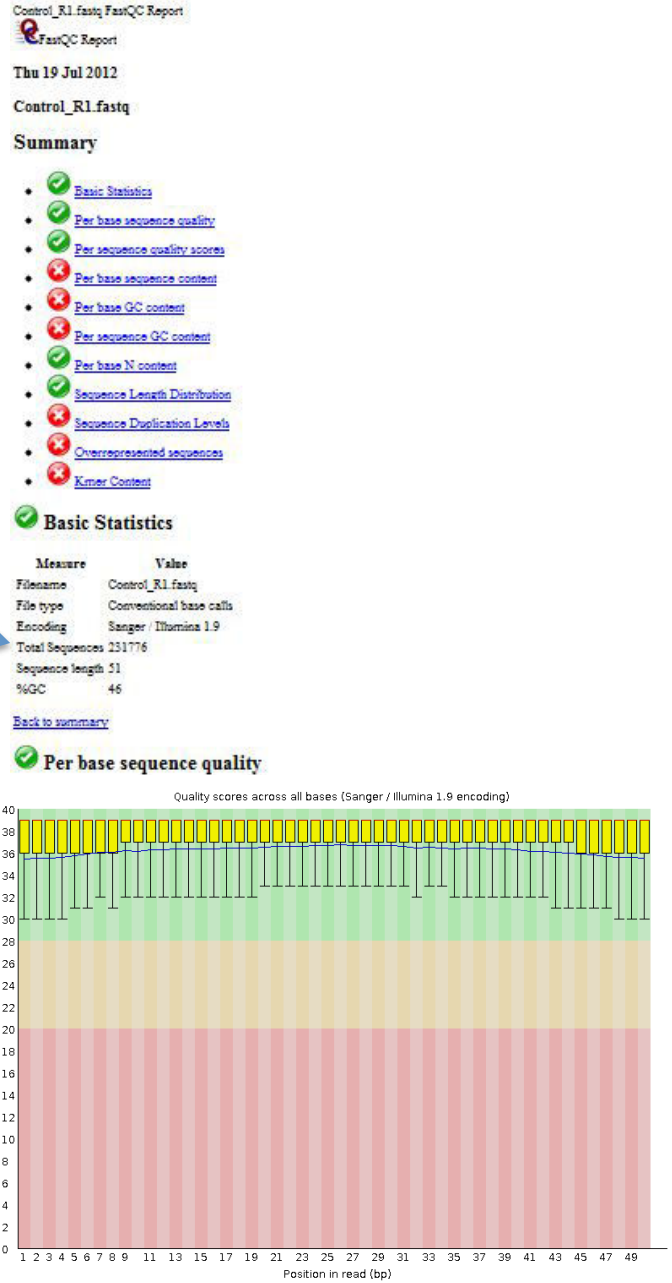
- Tools > ⚙ > Tool Search
- Enter “fastqc”
- Select the tool
- Enter a data set & Execute
- Trick for quick entry of multiple samples:
  - Immediately hit “BACK”
  - Update parameters and re-execute



FastQC Output Report

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

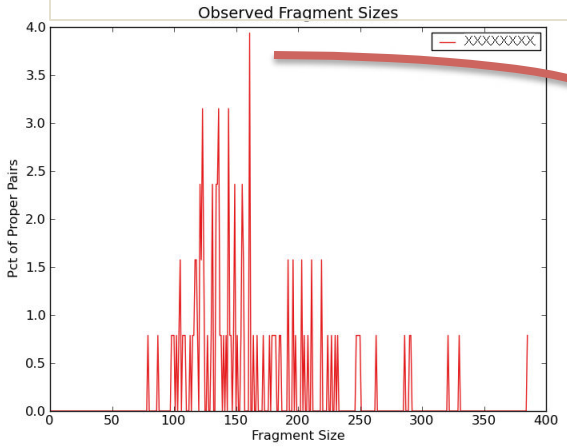


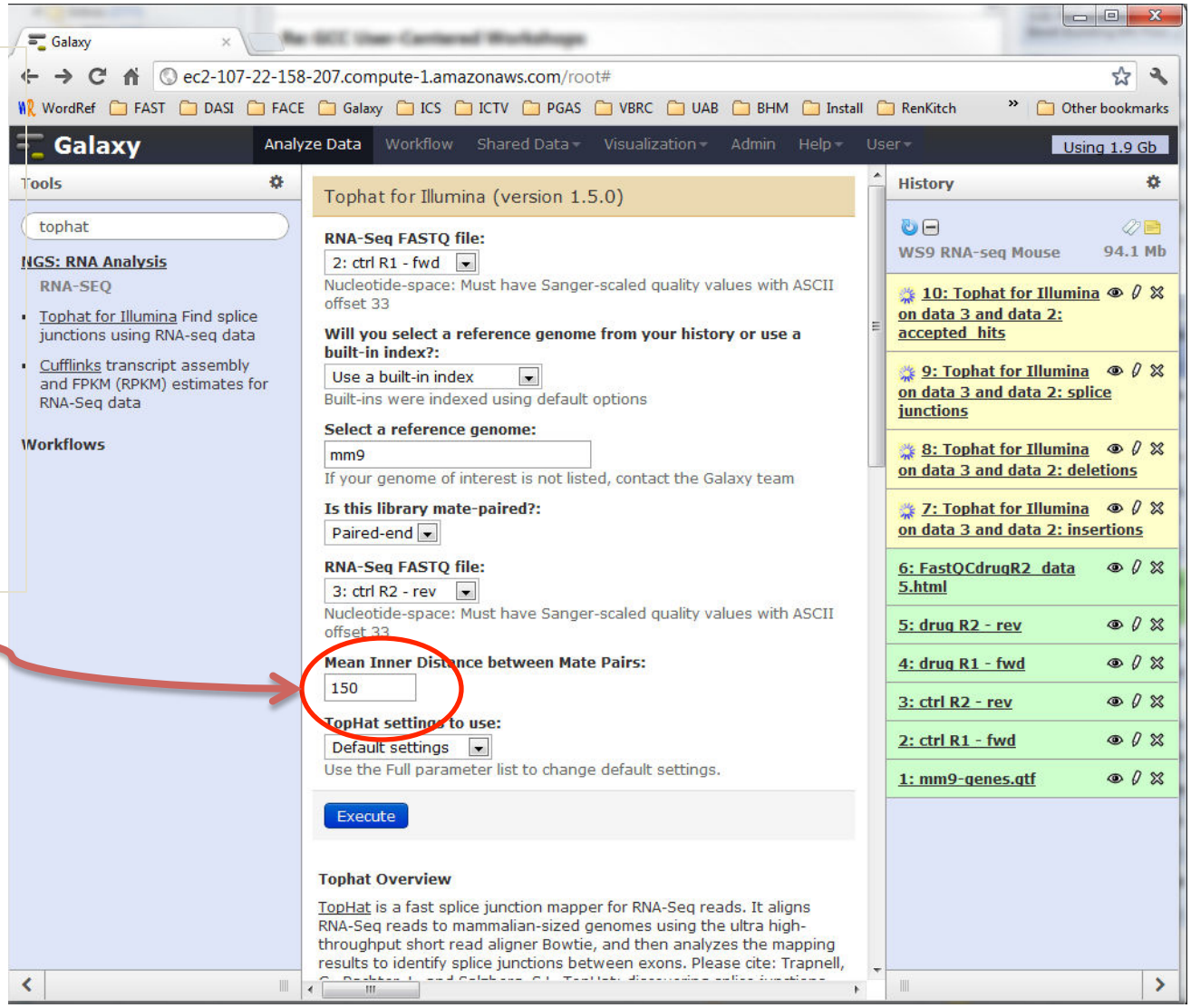


Tophat: map reads, create transcriptome

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)



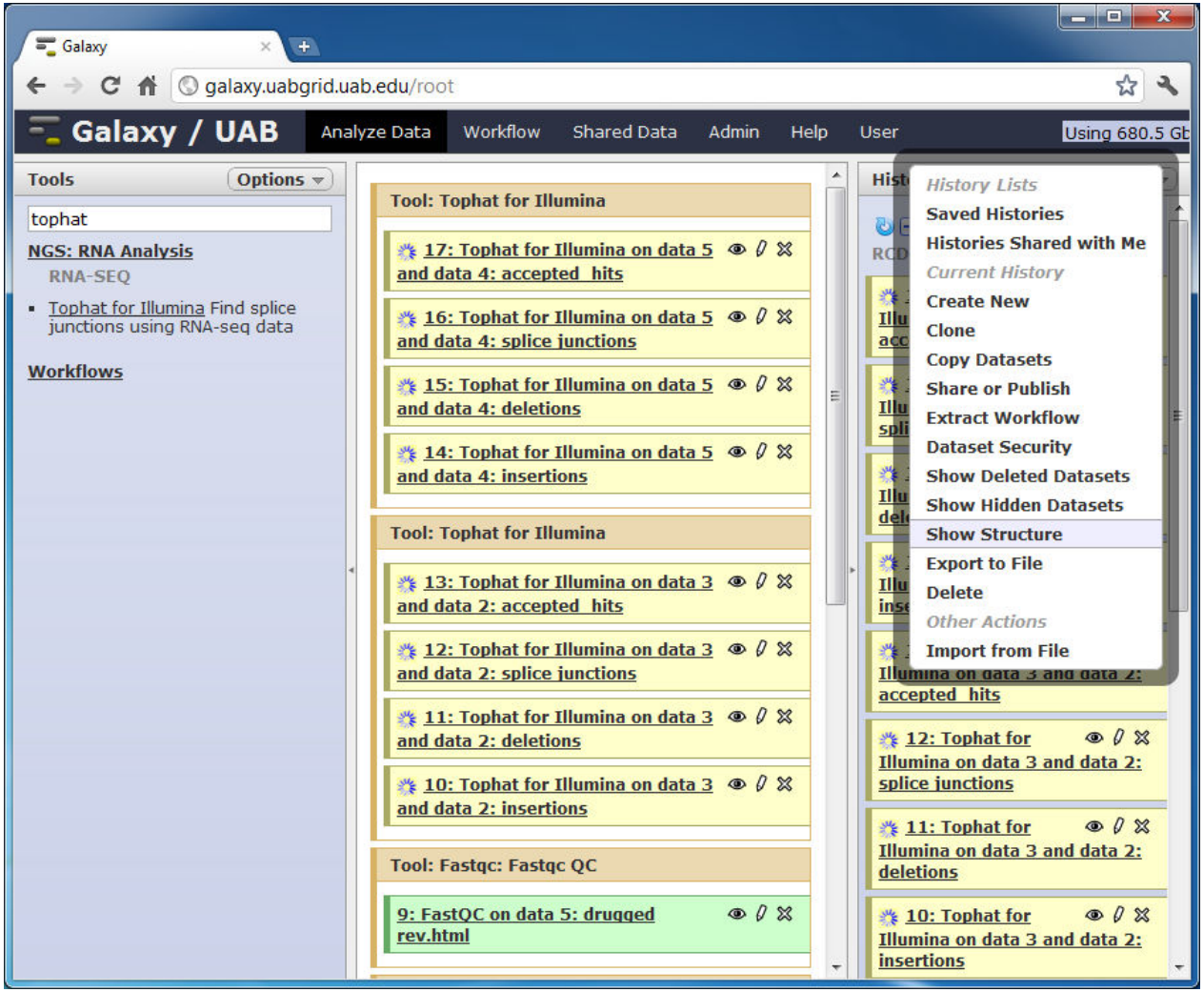


Launched 2 tophats (4 outputs each)

TIP:

> Show Structure groups datasets produced by the same tool.

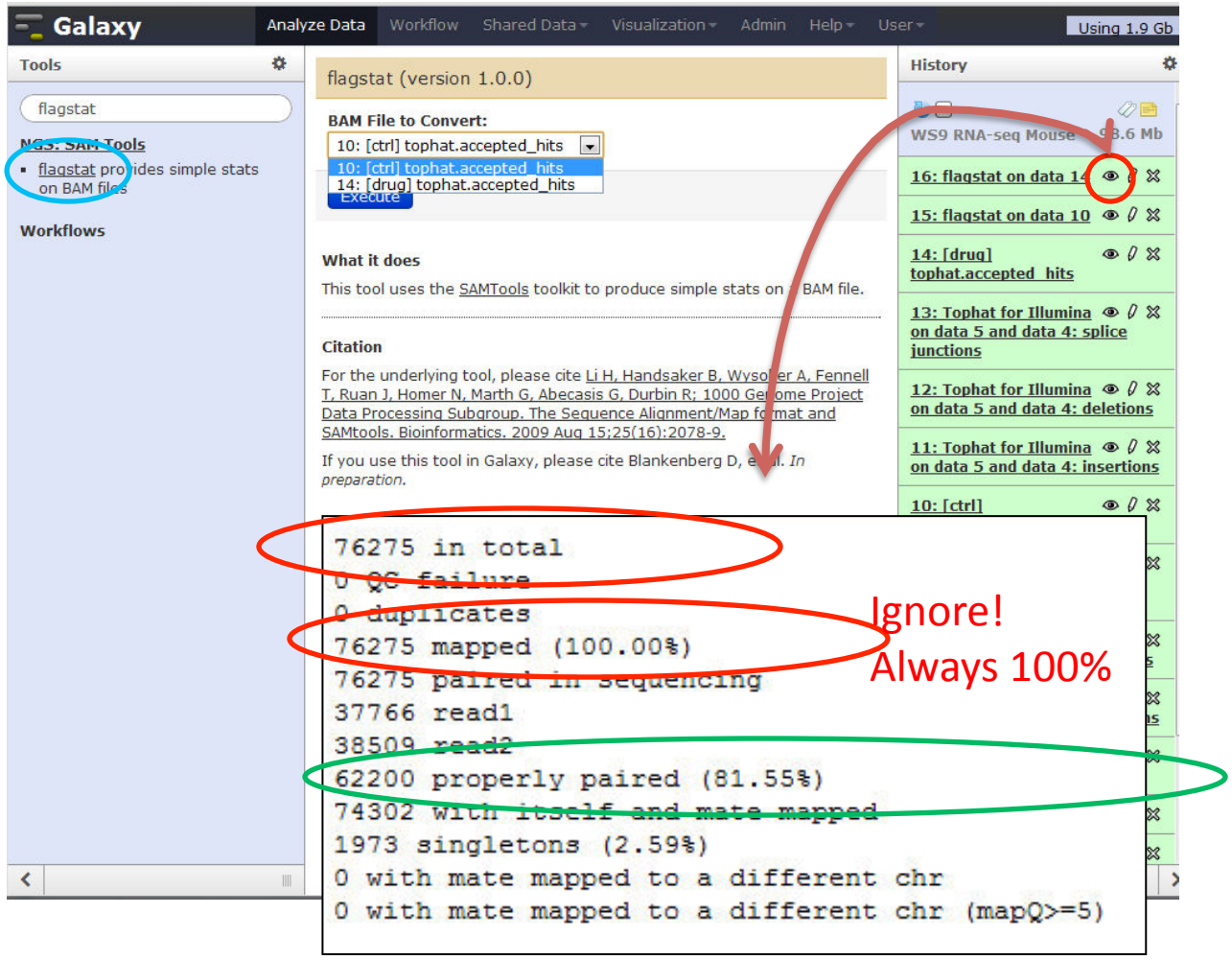
Edit “accepted_hits” Datasets to give shorter names with sample info: *[ctrl] tophat.accepted_hits*



Launch 2 flagstats: QC mapping

TIP:
You can execute a tool before it's predecessors have finished running.

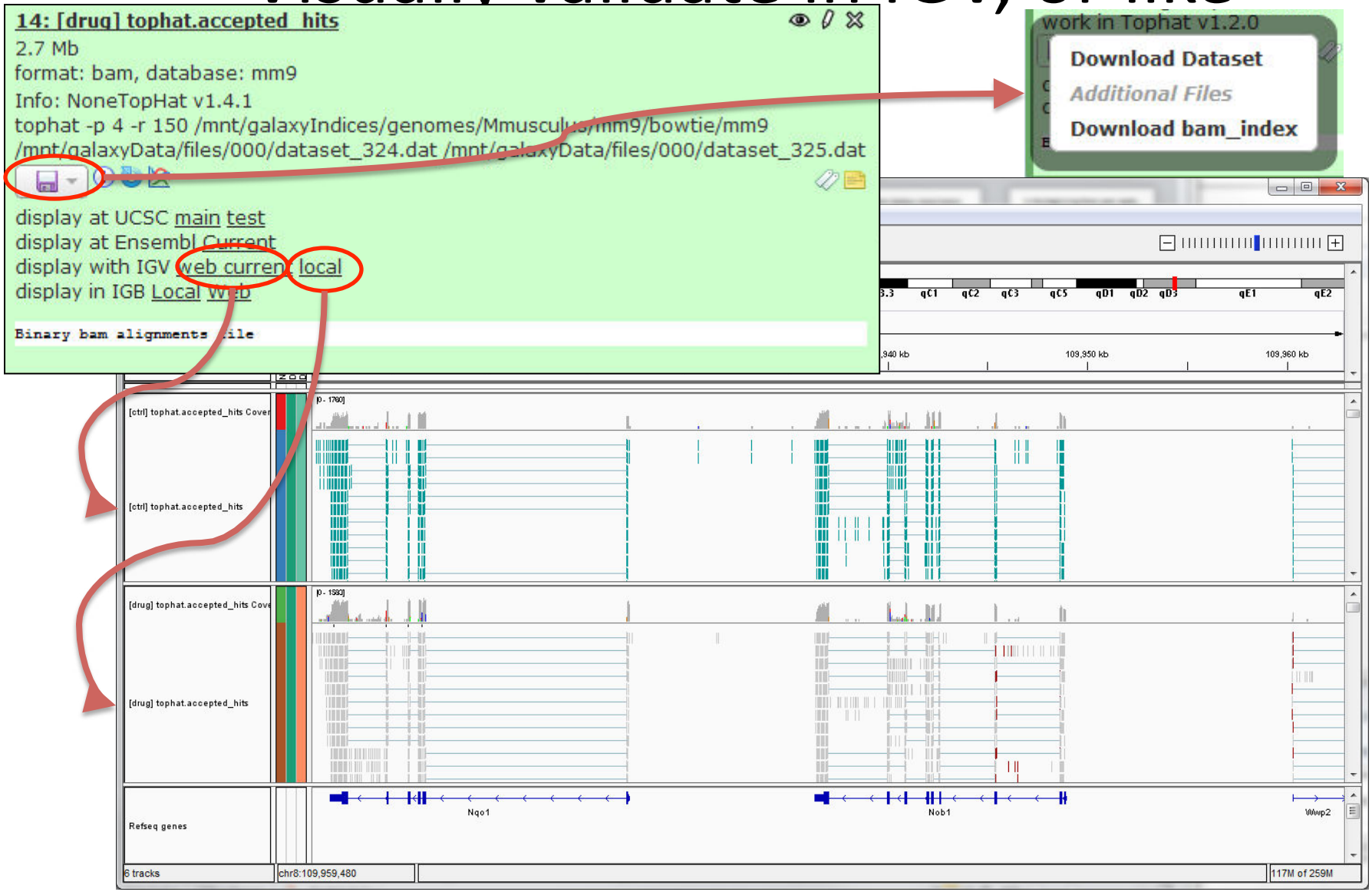
Galaxy will queue it until all the dependencies are satisfied!



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

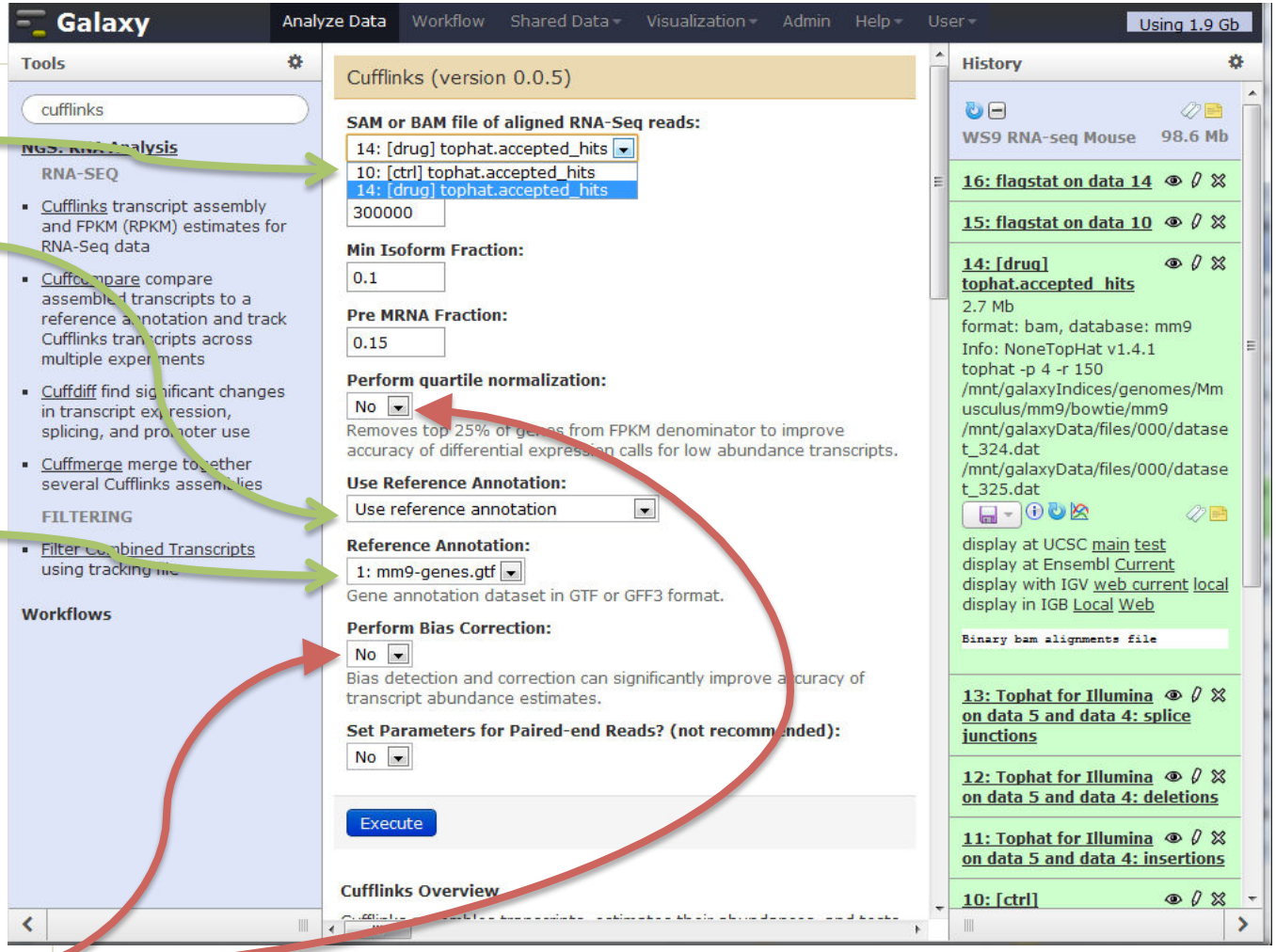
Read data:
accepted_hits (BAM)

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

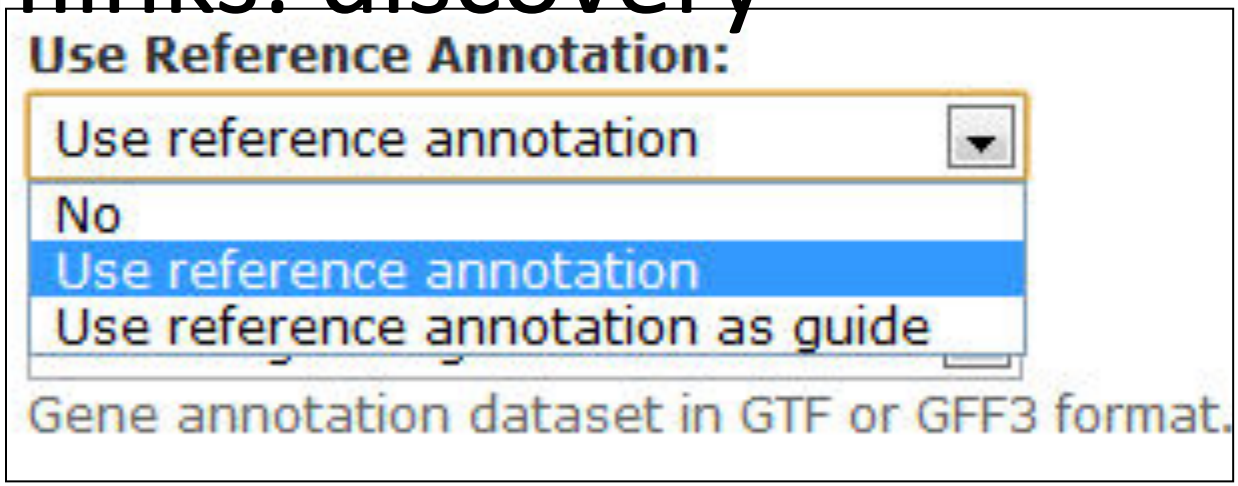
Reference annotation:
iGenome genes.GTF

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



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

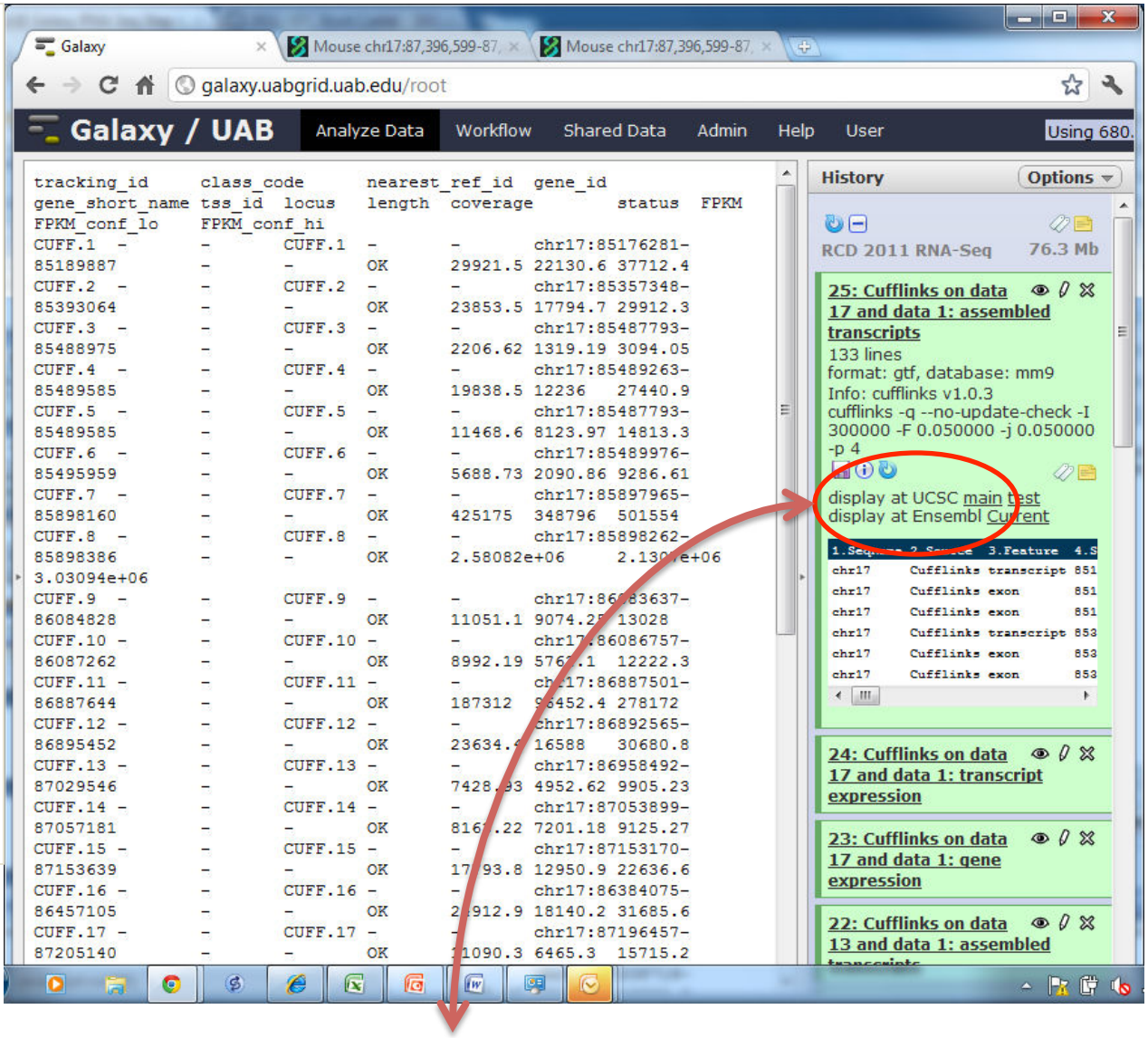
Cufflinks outputs

TIP: hide the cmd bar!

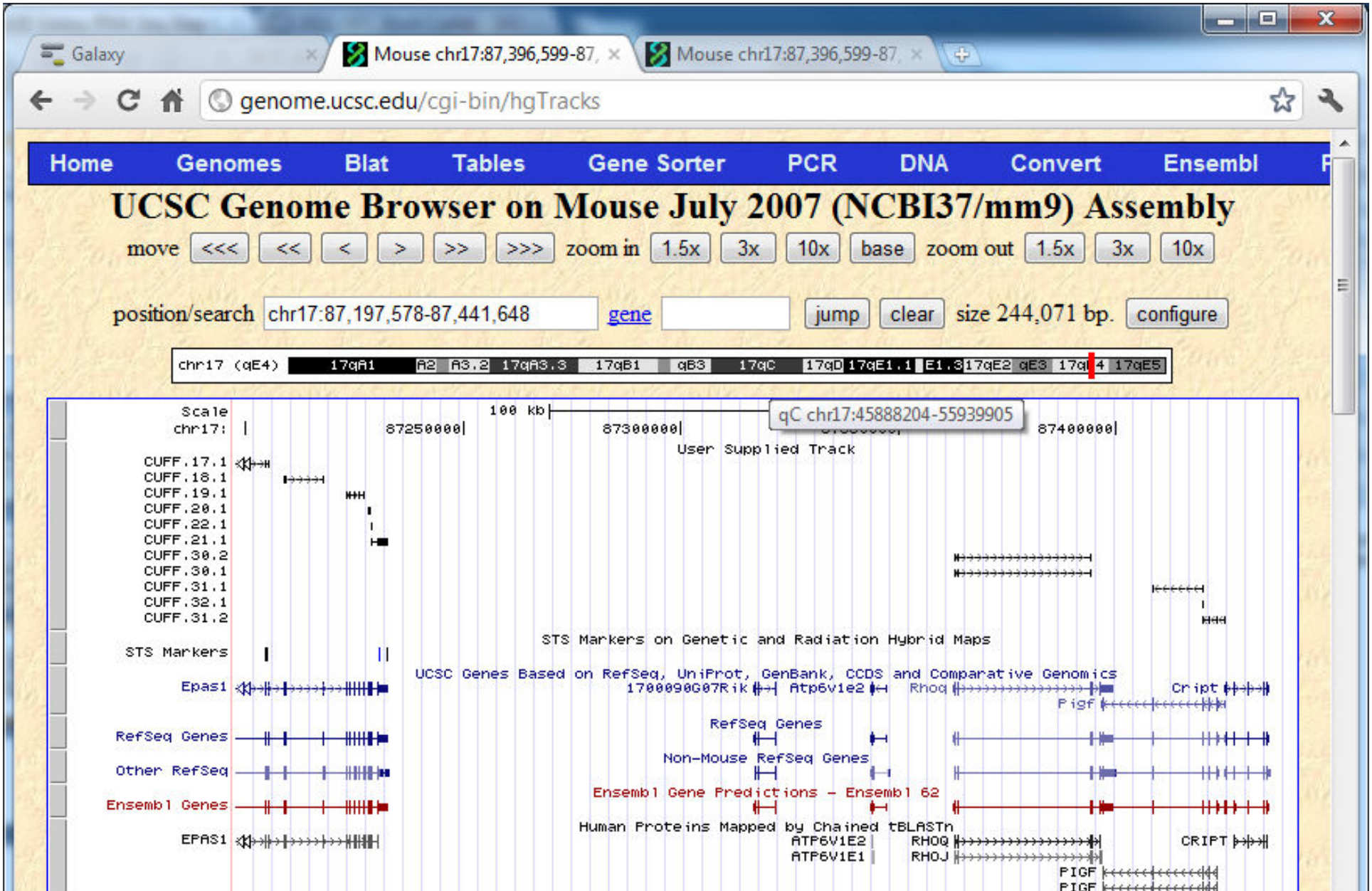
assembled_transcripts (GTF)
list of isoforms
transcript_expression (tab)
isoforms w/ FPKM
gene_expression (tab)
genes w/ FPKM

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)



assembled_transcripts Visualization

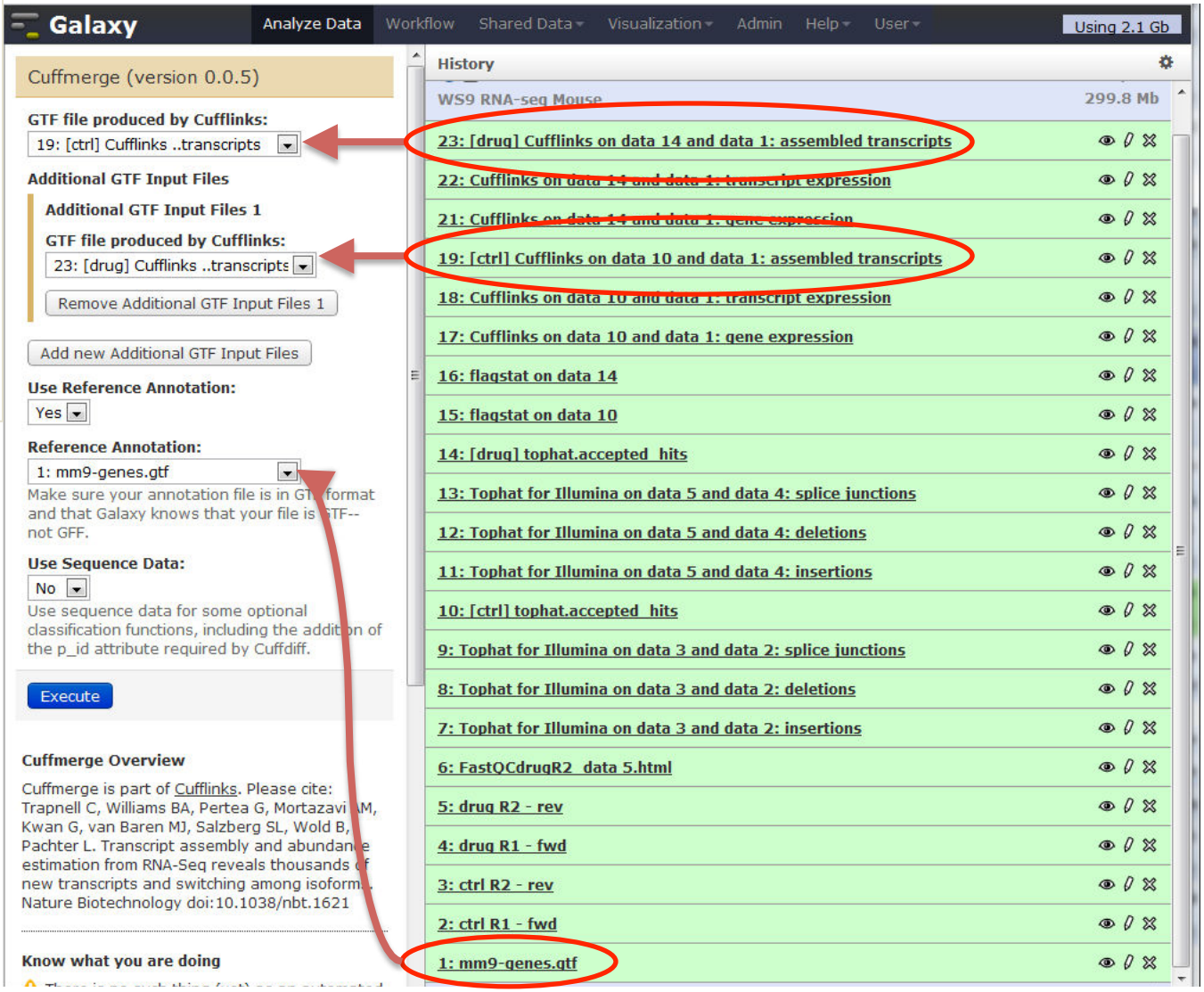


Cuffmerge: combine transcripts

samples & replicates -> complete transcriptome

Run once, using both samples

Inputs
[ctrl] assembled _trans…
[drug] assembled_trans…
Reference_annotation (GTF)



Cuffdiff: Quantize Transcripts

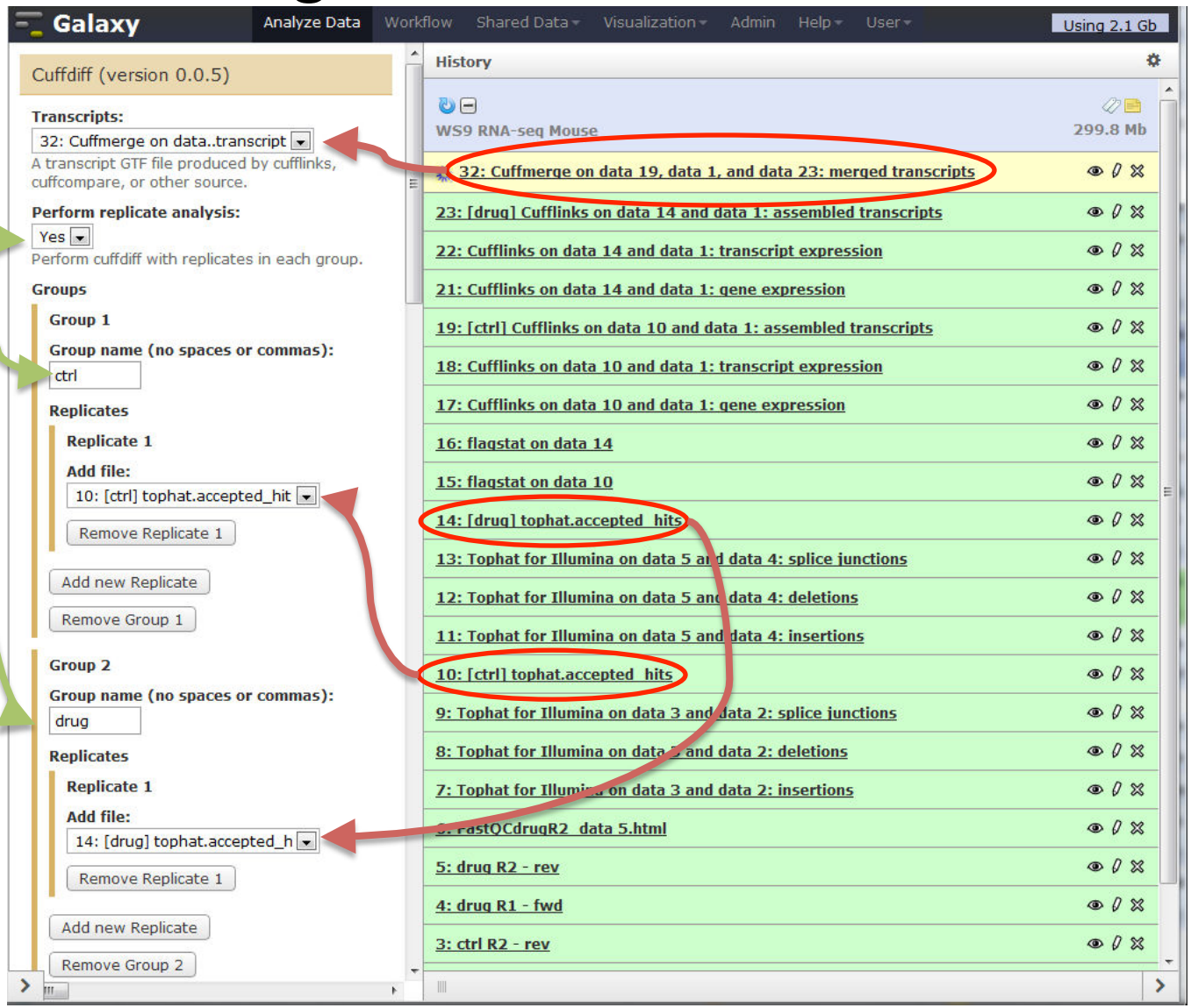
Compute fold change between conditions

Run once,
use all samples & replicates

ALWAYS create replicate groups for each condition

ALWAYS name your groups

even if you have only one replicate!

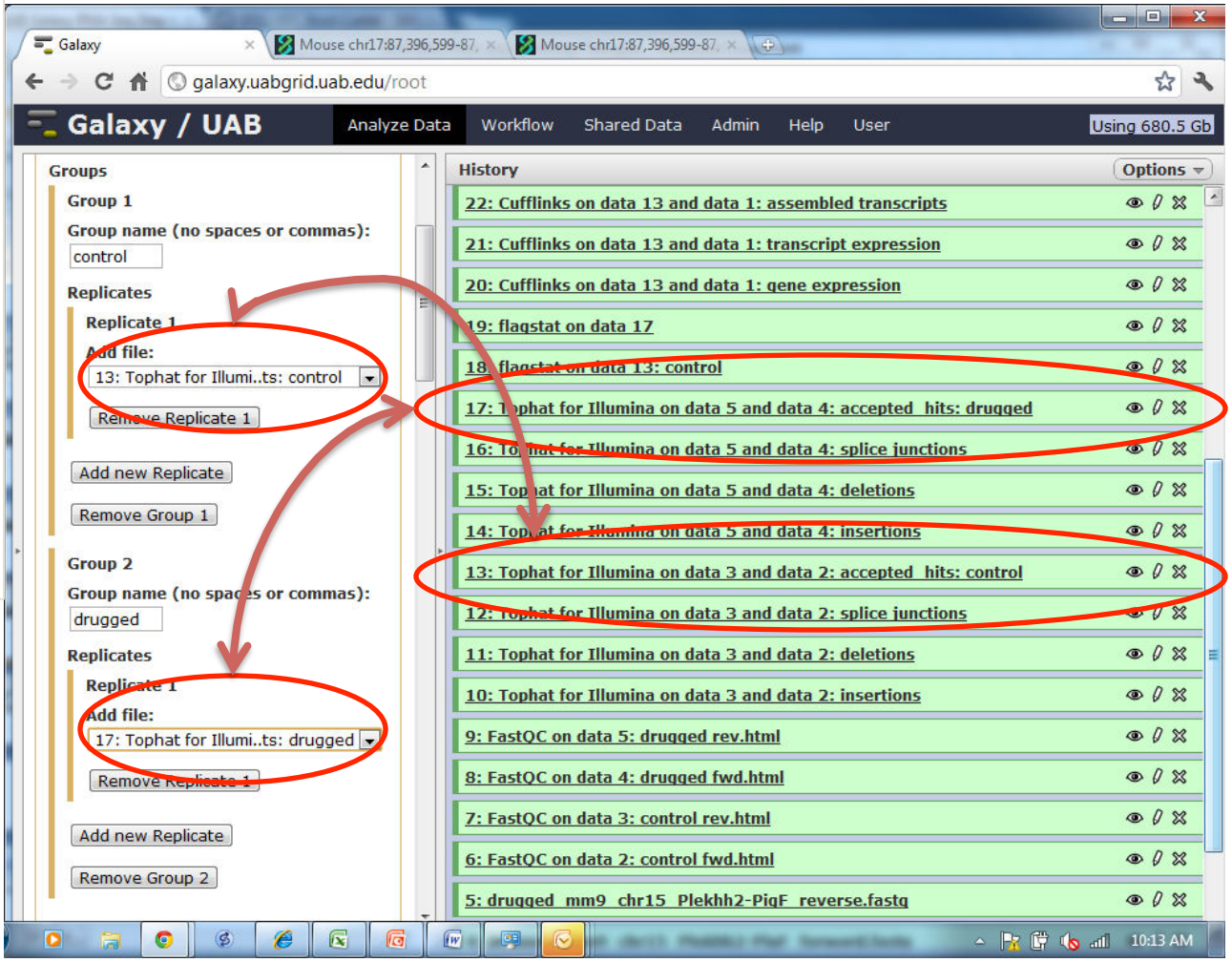


Cuffdiff: Fold Change Between Conditions

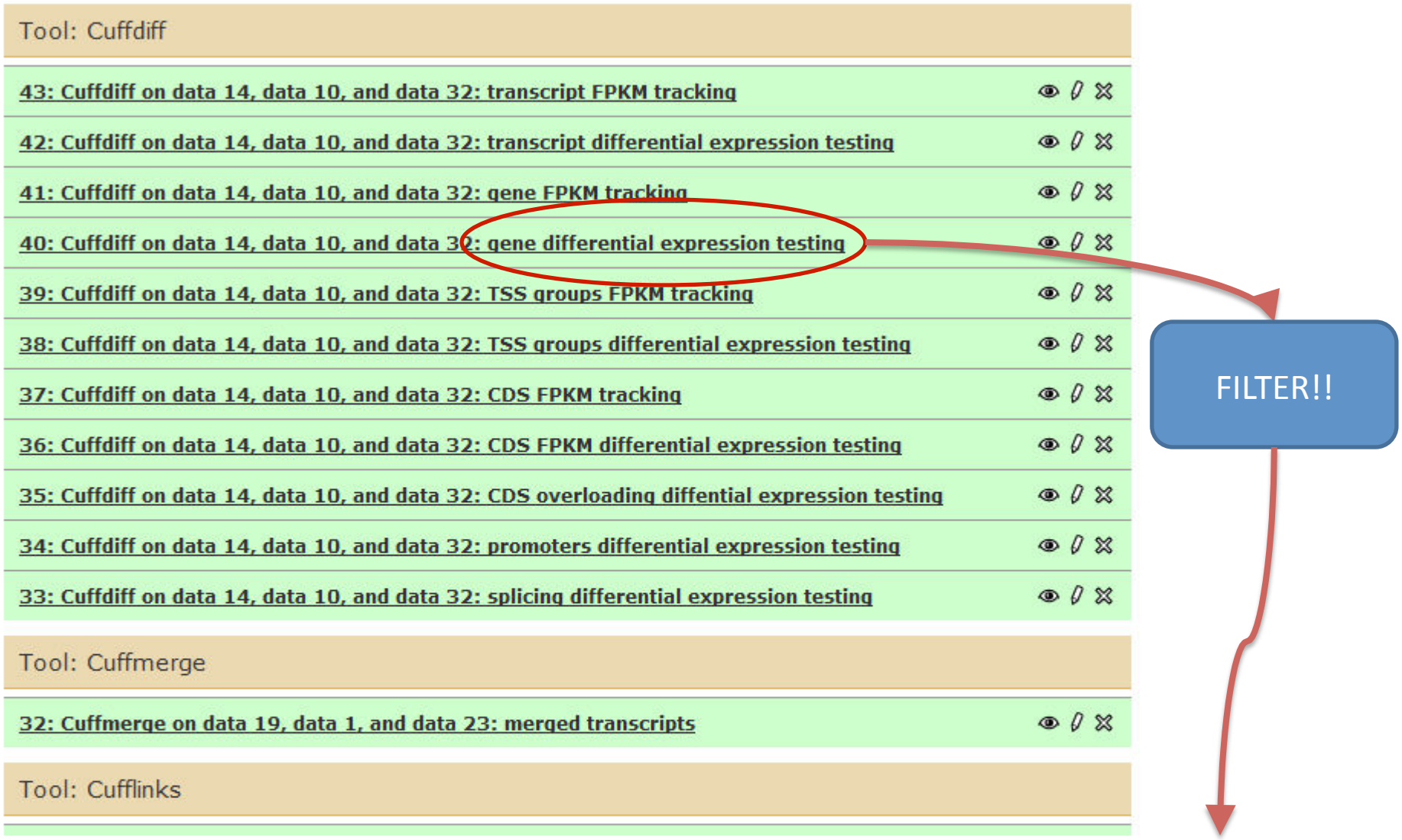
Run once, using both samples

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

Short names – no spaces or special characters!!!

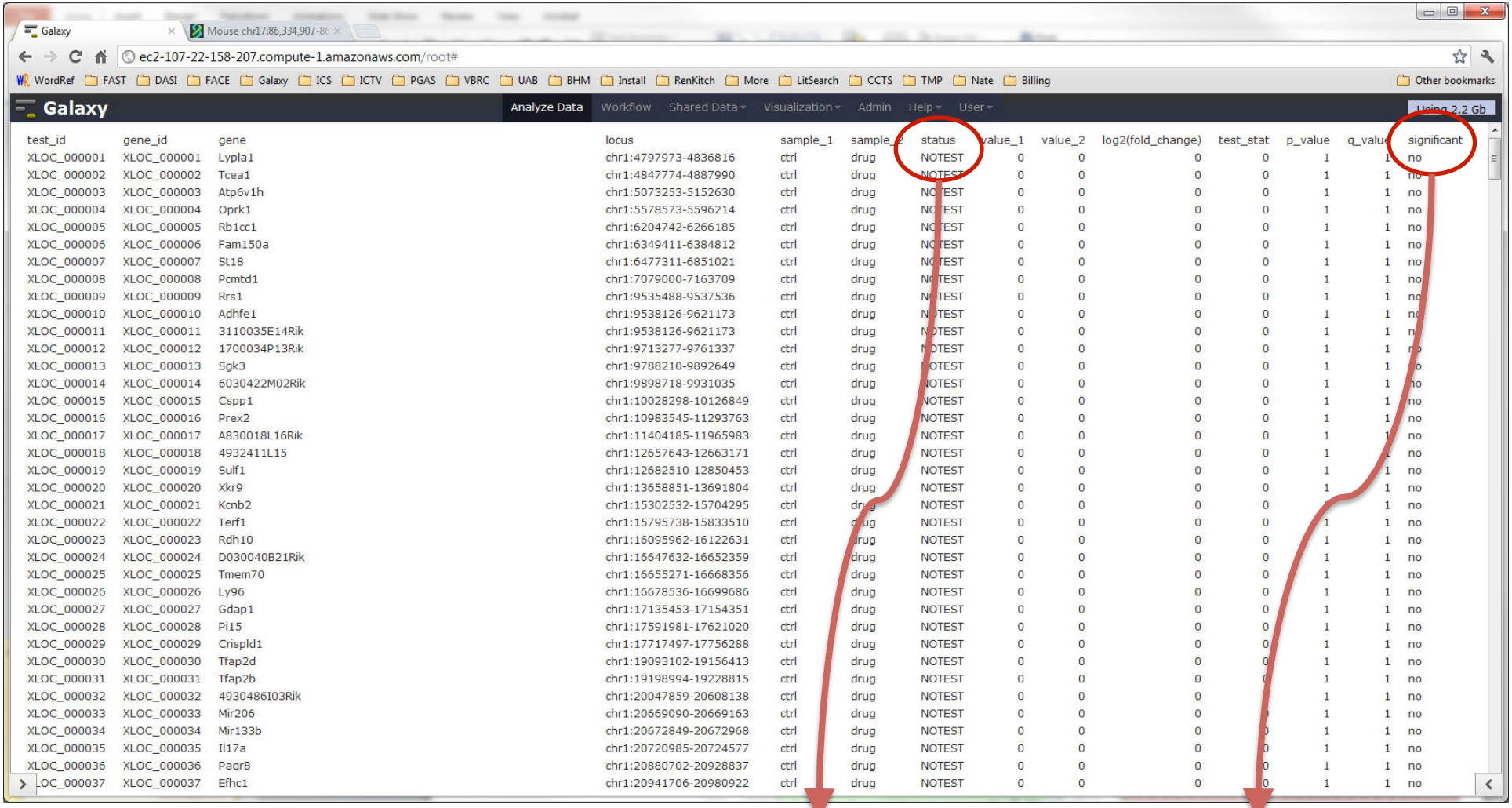


Cuffdiff output
(> view structure)



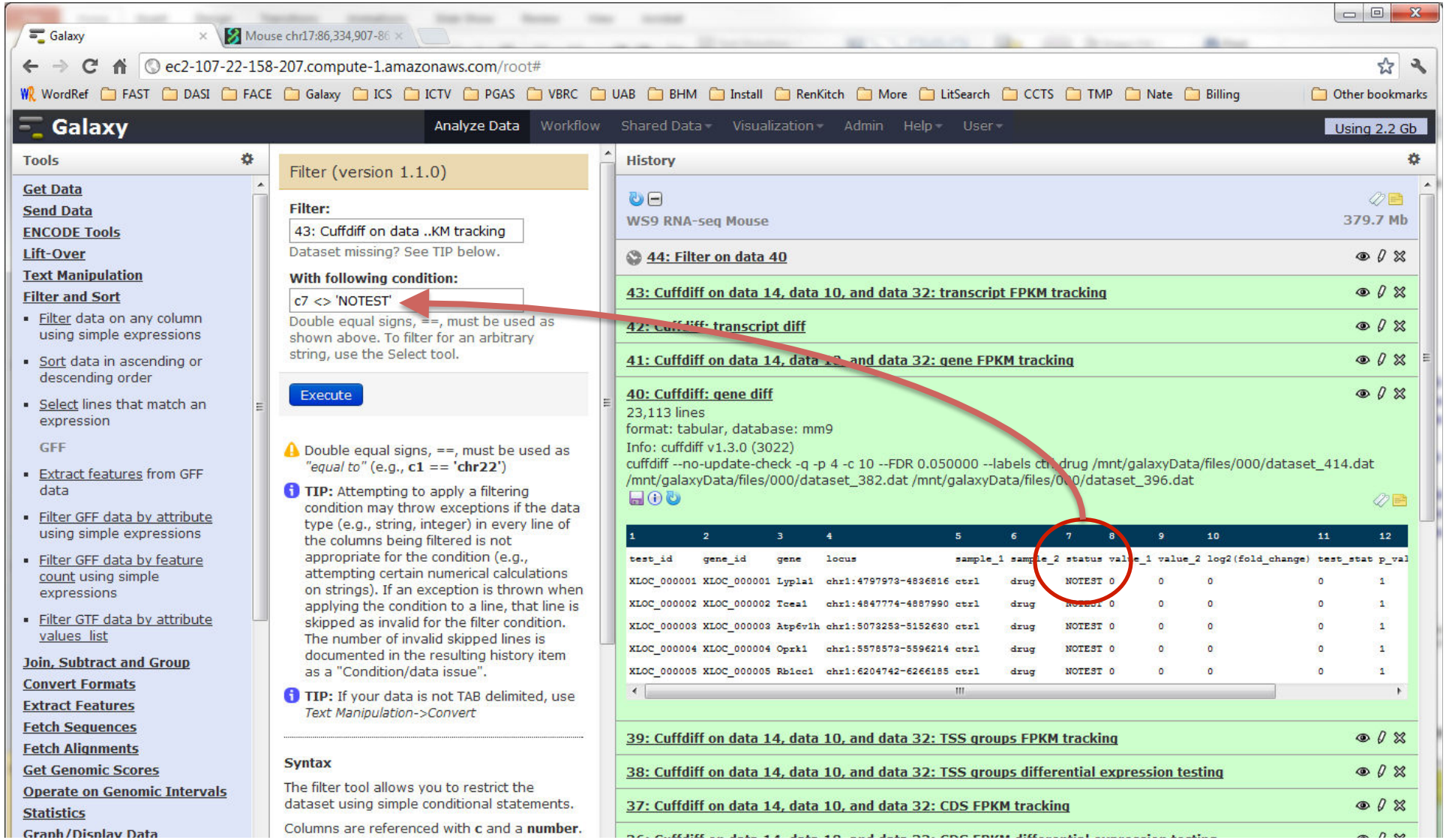
Cuffdiff “gene diff”: 23,113 lines!

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

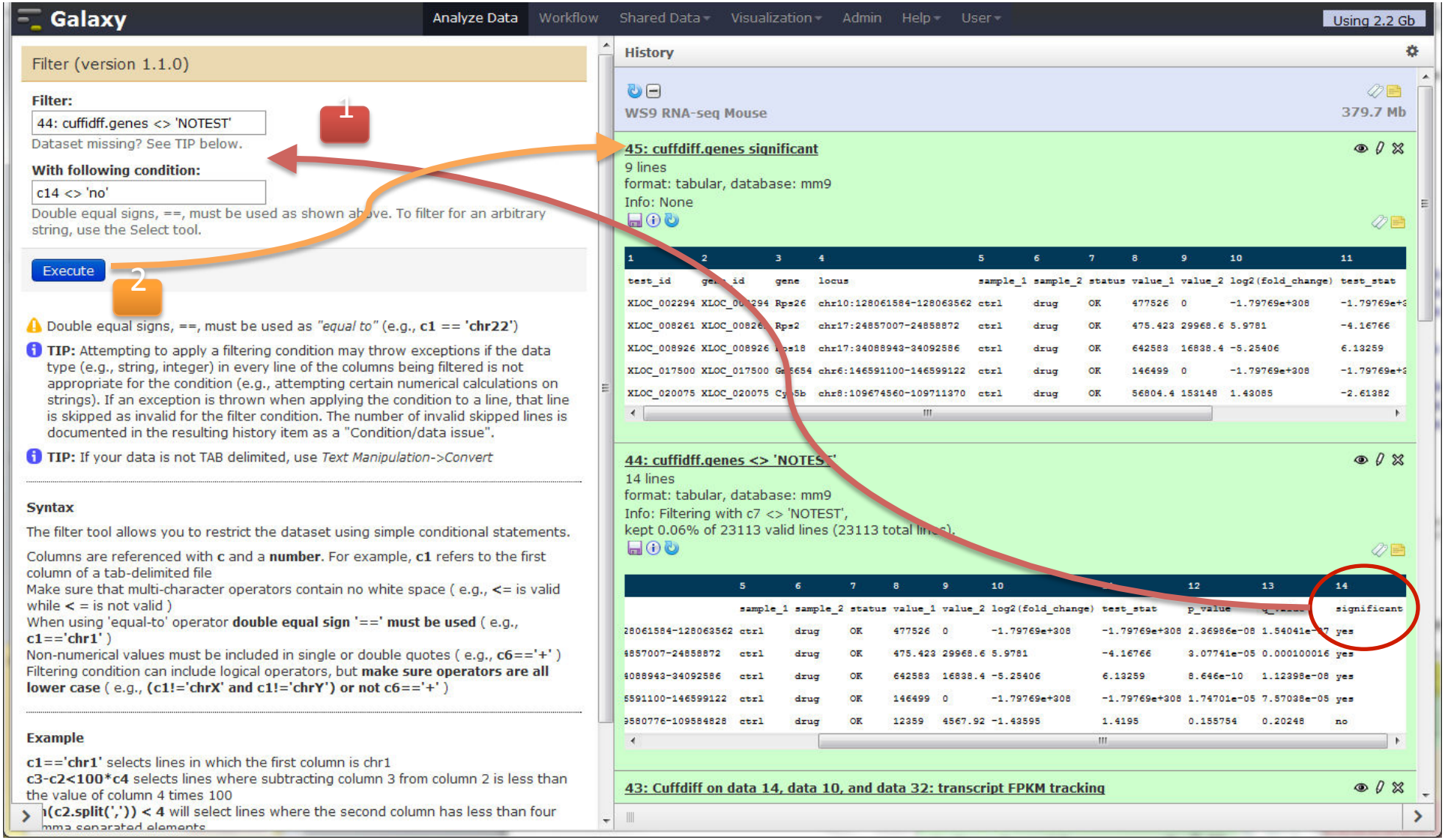


test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
XLOC_000001	XLOC_000001	Lypla1	chr1:4797973-4836816	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000002	XLOC_000002	Tcea1	chr1:4847774-4887990	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000003	XLOC_000003	Atp6v1h	chr1:5073253-5152630	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000004	XLOC_000004	Oprk1	chr1:5578573-5596214	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000005	XLOC_000005	Rb1cc1	chr1:6204742-6266185	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000006	XLOC_000006	Fam150a	chr1:6349411-6384812	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000007	XLOC_000007	St18	chr1:6477311-6851021	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000008	XLOC_000008	Pcmtd1	chr1:7079000-7163709	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000009	XLOC_000009	Rrs1	chr1:9535488-9537536	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000010	XLOC_000010	Adhfe1	chr1:9538126-9621173	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000011	XLOC_000011	3110035E14Rik	chr1:9538126-9621173	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000012	XLOC_000012	1700034P13Rik	chr1:9713277-9761337	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000013	XLOC_000013	Sgk3	chr1:9788210-9892649	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000014	XLOC_000014	6030422M02Rik	chr1:9898718-9931035	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000015	XLOC_000015	Cspp1	chr1:10028298-10126849	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000016	XLOC_000016	Prex2	chr1:10983545-11293763	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000017	XLOC_000017	A830018L16Rik	chr1:11404185-11965983	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000018	XLOC_000018	4932411L15	chr1:12657643-12663171	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000019	XLOC_000019	Sulf1	chr1:12682510-12850453	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000020	XLOC_000020	Xkr9	chr1:13658851-13691804	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000021	XLOC_000021	Kcnb2	chr1:15302532-15704295	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000022	XLOC_000022	Terf1	chr1:15795738-15833510	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000023	XLOC_000023	Rdh10	chr1:16095962-16122631	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000024	XLOC_000024	D030040B21Rik	chr1:16647632-16652359	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000025	XLOC_000025	Tmem70	chr1:16655271-16668356	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000026	XLOC_000026	Ly96	chr1:16678536-16699686	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000027	XLOC_000027	Gdap1	chr1:17135453-17154351	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000028	XLOC_000028	Pi15	chr1:17591981-17621020	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000029	XLOC_000029	Crispld1	chr1:17717497-17756288	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000030	XLOC_000030	Tfap2d	chr1:19093102-19156413	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000031	XLOC_000031	Tfap2b	chr1:19198994-19228815	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000032	XLOC_000032	4930486I03Rik	chr1:20047859-20608138	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000033	XLOC_000033	Mir206	chr1:20669090-20669163	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000034	XLOC_000034	Mir133b	chr1:20672849-20672968	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000035	XLOC_000035	Il17a	chr1:20720985-20724577	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000036	XLOC_000036	Paqr8	chr1:20880702-20928837	ctrl	drug	NOTEST	0	0	0	0	1	1	no
XLOC_000037	XLOC_000037	Efhc1	chr1:20941706-20980922	ctrl	drug	NOTEST	0	0	0	0	1	1	no

Gene diff: genes with enough reads
14 genes



Gene diff: statistically significant
9 genes



“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	significar
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

Galaxy37-[genes_exp____NOTEST_].tabular - Microsoft Excel

File Home Insert Page Layout Formulas Data Review View

H16

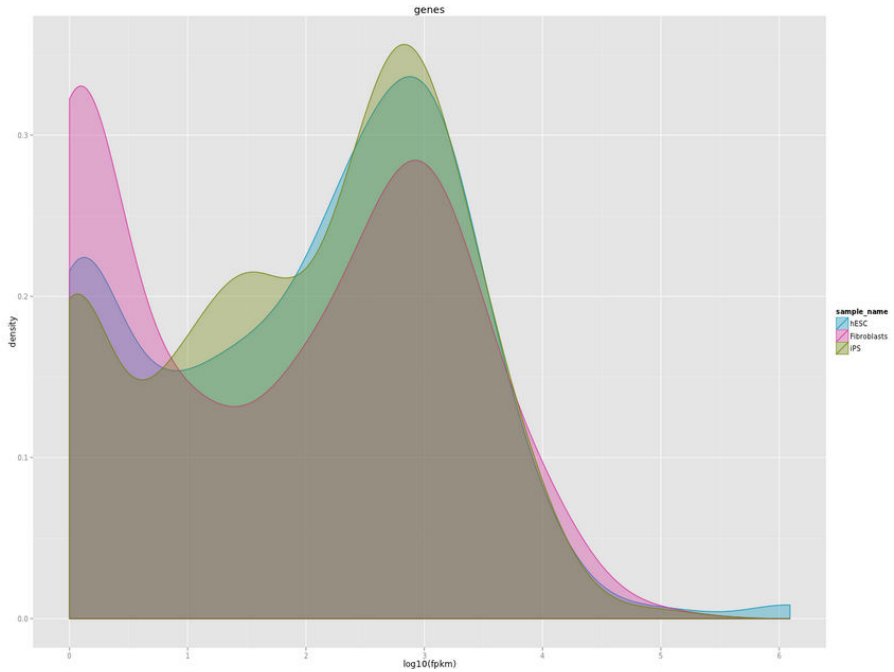
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
2	XLOC_008261	XLOC_008261	Rps2	chr17:24857007-24858872	ctrl	drug	OK	12,155	20,347	0.743285	-1.3184	0.1874	0.2186	no
3	XLOC_020075	XLOC_020075	Cyb5b	chr8:109674560-109711370	ctrl	drug	OK	73,935	103,978	0.491939	-1.9941	0.0461	0.0646	no
4	XLOC_020077	XLOC_020077	Wwp2	chr8:109960297-110082495	ctrl	drug	OK	39,072	3,288	-3.571	7.5735	0.0000	0.0000	yes
5	XLOC_020591	XLOC_020591	Terf2	chr8:109593301-109620489	ctrl	drug	OK	68,292	106,354	0.639088	-3.1602	0.0016	0.0028	yes
6	XLOC_020592	XLOC_020592	Nqo1	chr8:109912124-109927105	ctrl	drug	OK	-	14,465			0.0000	0.0001	yes
7	XLOC_020594	XLOC_020594	Psmd7	chr8:110104279-110112382	ctrl	drug	OK	54,160	23,571	-1.20026	3.5457	0.0004	0.0009	yes
8	XLOC_020696	XLOC_020696	-	chr8:109786639-109787183	ctrl	drug	OK	876,105	907,362	0.0505752	-0.2017	0.8401	0.8401	no
9														

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)

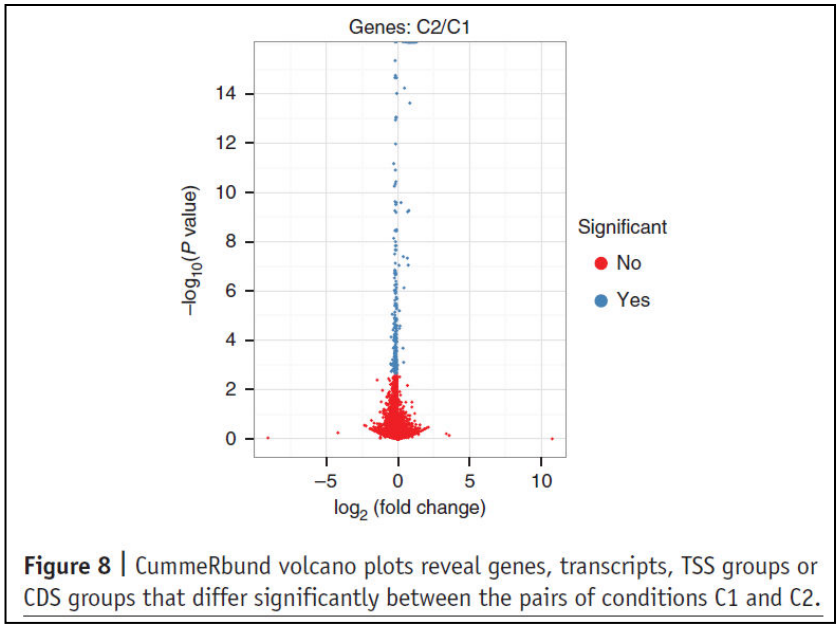


Figure 8 | CummeRbund volcano plots reveal genes, transcripts, TSS groups or CDS groups that differ significantly between the pairs of conditions C1 and C2.

(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