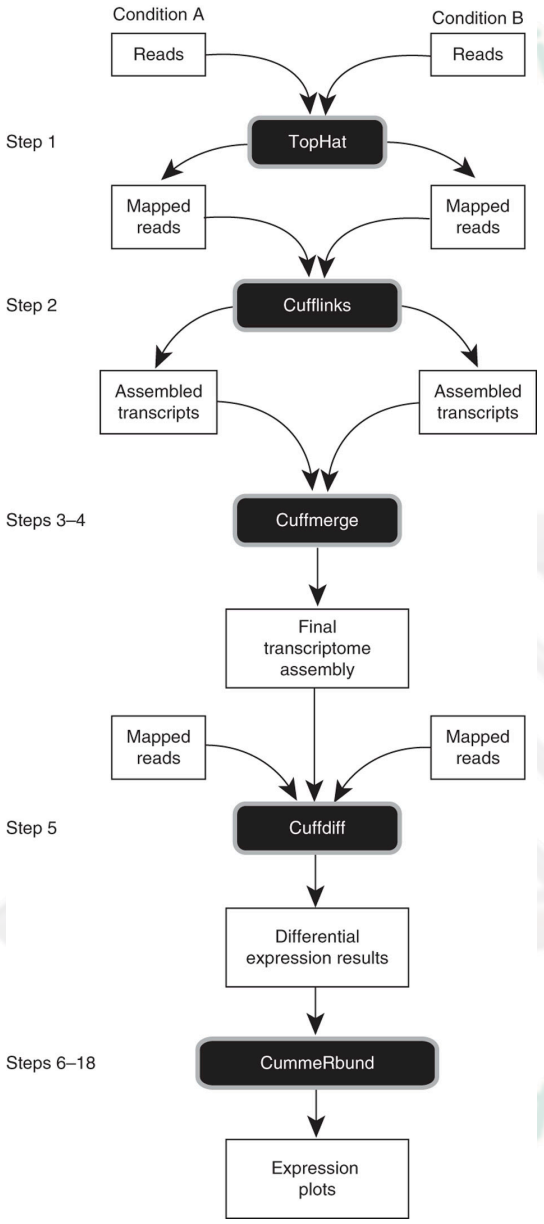
Transcriptome and Pathway Analysis

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

Immersion Course

RNA-Seq pipeline



Bowtie
Extremely fast, general purpose short read aligner

TopHat
Aligns RNA-Seq reads to the genome using Bowtie
Discovers splice sites

Cufflinks package

Cufflinks
Assembles transcripts

Cuffcompare
Compares transcript assemblies to annotation

Cuffmerge
Merges two or more transcript assemblies

Cuffdiff
Finds differentially expressed genes and transcripts
Detects differential splicing and promoter use

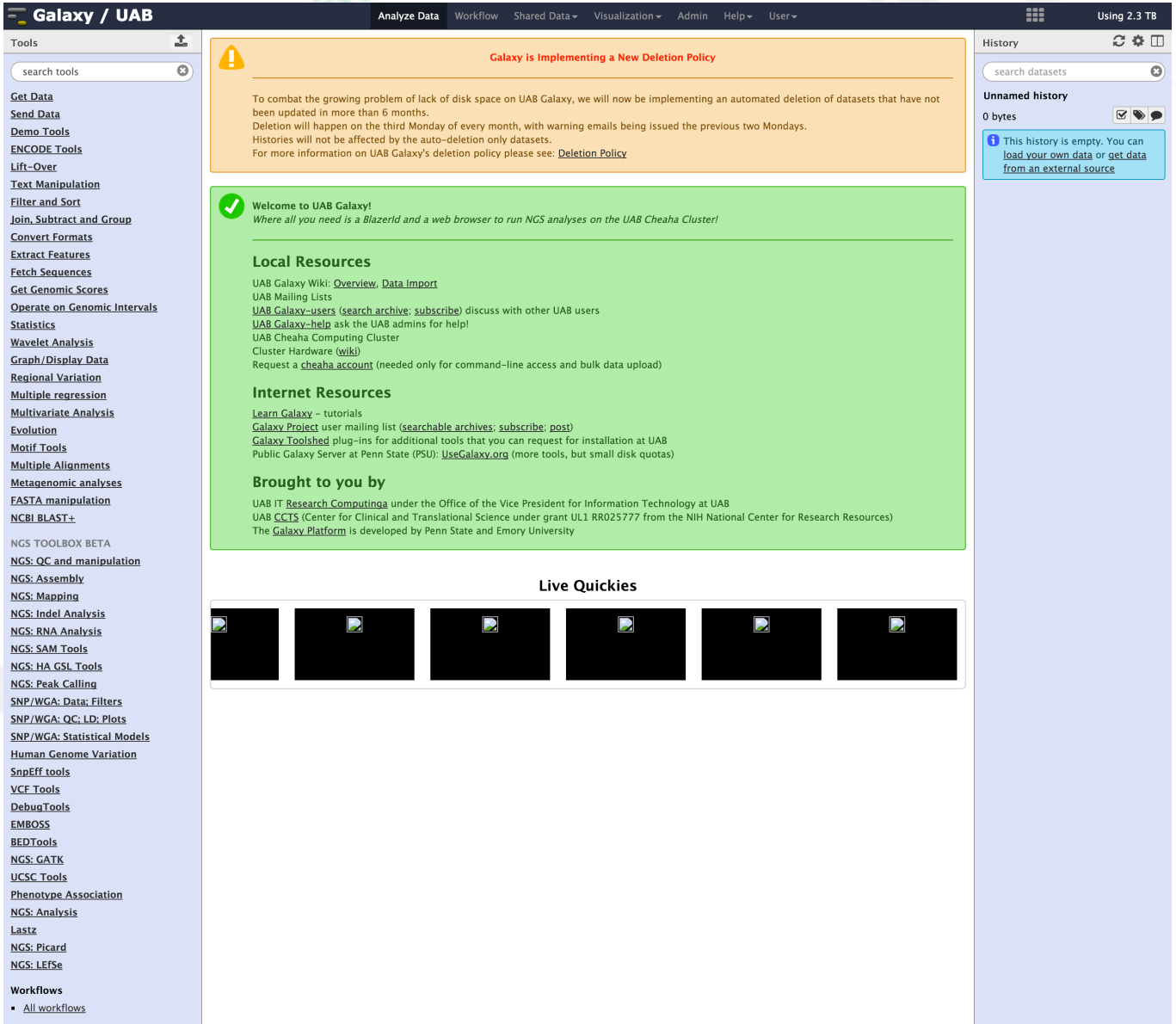
CummeRbund
Plots abundance and differential expression results from Cuffdiff

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

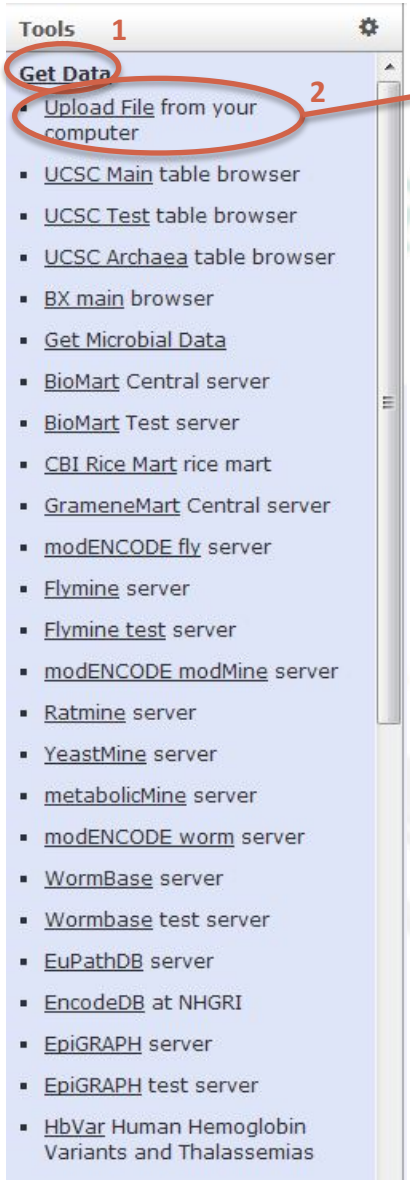
Galaxy Splash Page

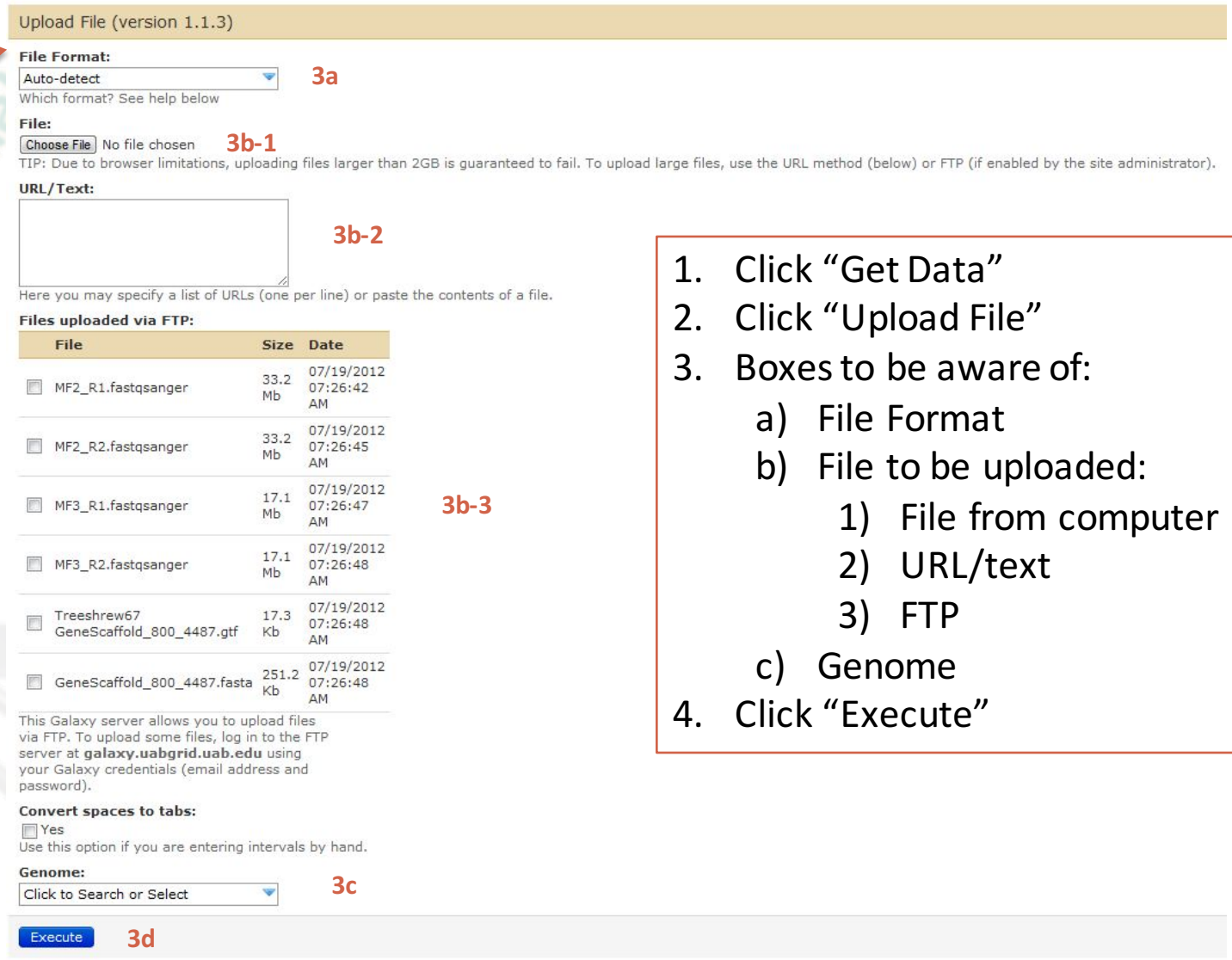
https://www.uab.edu/galaxy

https://main.g2.bx.psu.edu/



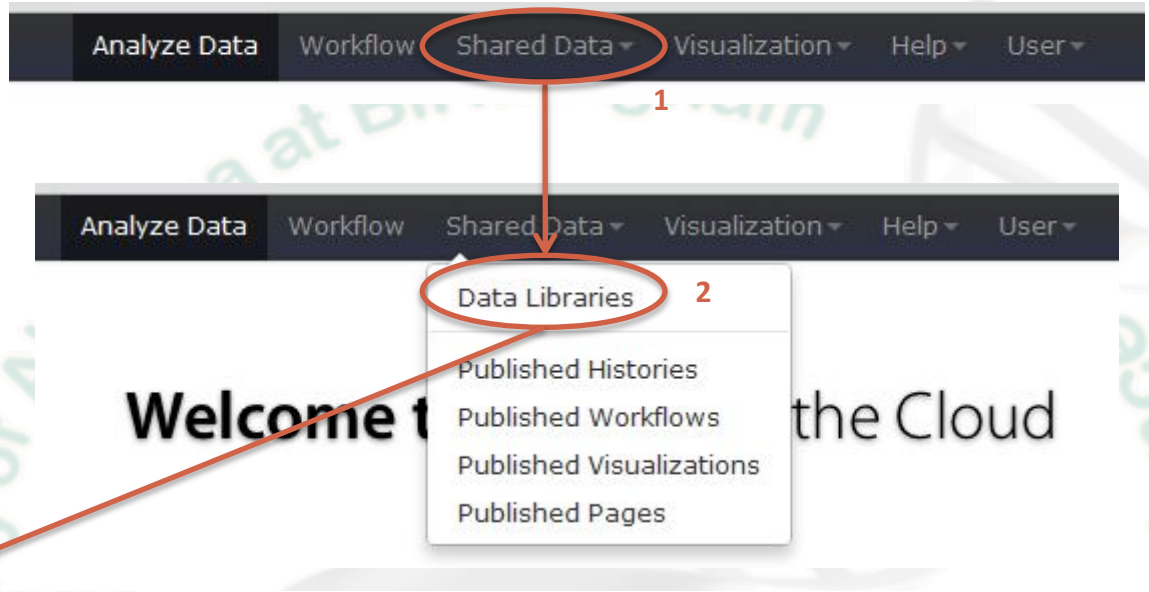
Upload/Import Data





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 "Immersion course prep"

Name	Message	Data type	Date uploaded	File size
Control_rep1_r1.fastq		fastqsanger	2012-08-06	14.4 Mb
Control_rep1_r2.fastq		fastqsanger	2012-08-06	14.4 Mb
Control_rep2_r1.fastq		fastqsanger	2012-08-06	14.4 Mb
Control_rep2_r2.fastq		fastqsanger	2012-08-06	14.4 Mb
Treated_rep1_r1.fastq		fastqsanger	2012-08-06	14.4 Mb
Treated_rep1_r2.fastq		fastqsanger	2012-08-06	14.4 Mb
Treated_rep2_r1.fastq		fastqsanger	2012-08-06	14.4 Mb
Treated_rep2_r2.fastq		fastqsanger	2012-08-06	14.4 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)

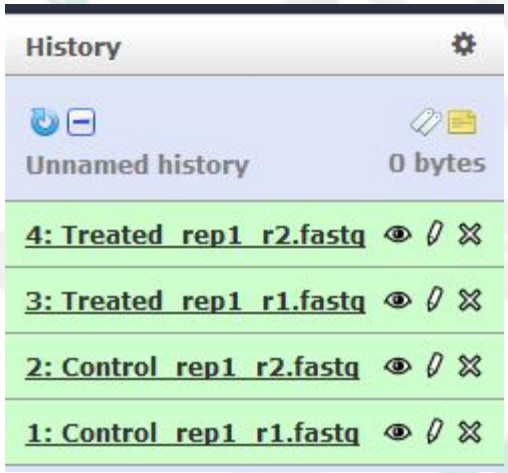
Import Shared Data to Current History

Data Library "Immersion course prep"

Name 1	Message	Data type	Date uploaded	File size
Control_rep1_r1.fastq		fastqsanger	2012-08-06	14.4 Mb
Control_rep1_r2.fastq		fastqsanger	2012-08-06	14.4 Mb
Control_rep2_r1.fastq		fastqsanger	2012-08-06	14.4 Mb
Control_rep2_r2.fastq		fastqsanger	2012-08-06	14.4 Mb
Treated_rep1_r1.fastq		fastqsanger	2012-08-06	14.4 Mb
Treated_rep1_r2.fastq		fastqsanger	2012-08-06	14.4 Mb
Treated_rep2_r1.fastq		fastqsanger	2012-08-06	14.4 Mb
Treated_rep2_r2.fastq		fastqsanger	2012-08-06	14.4 Mb

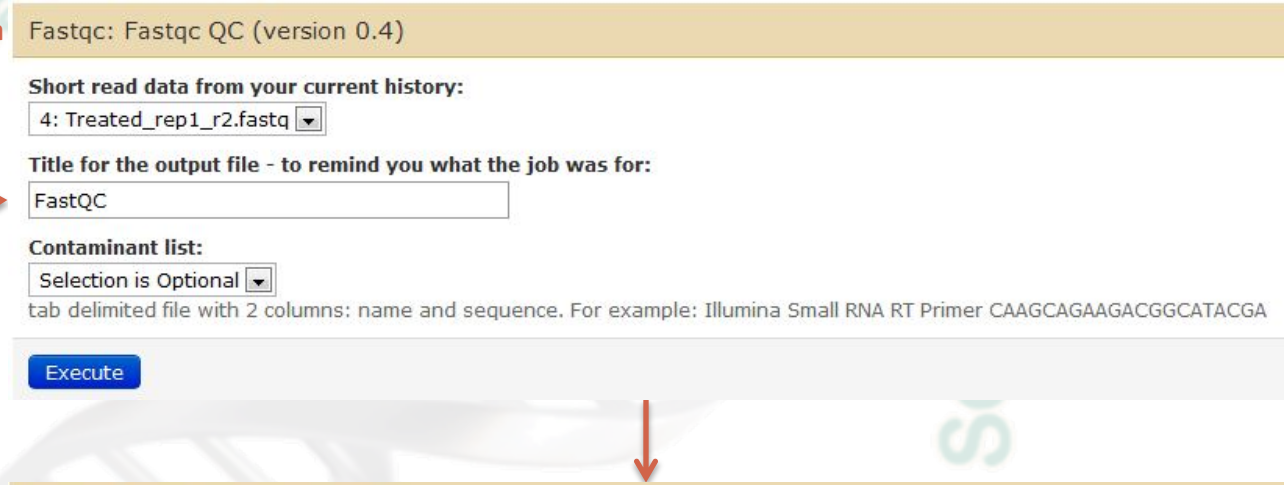
For selected datasets: Import to current history Go 2

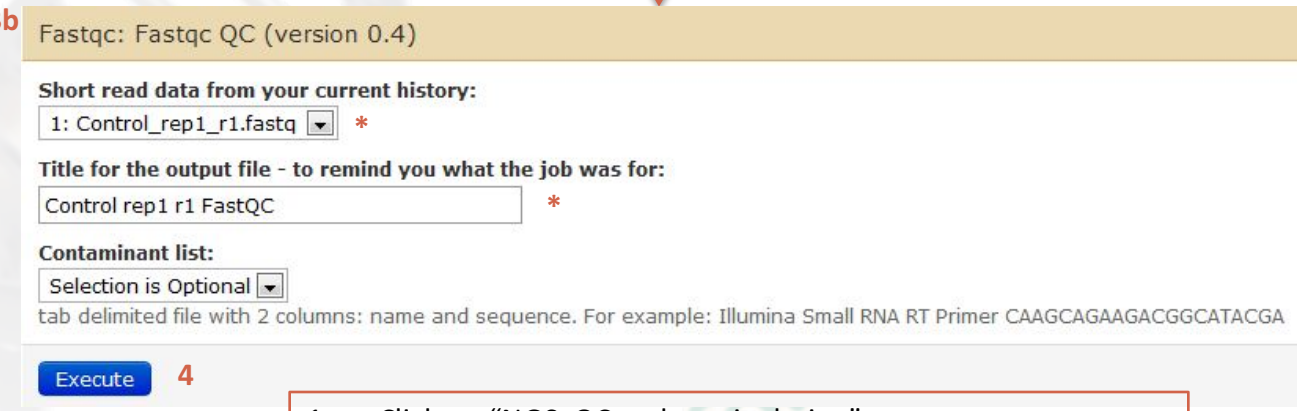
3



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

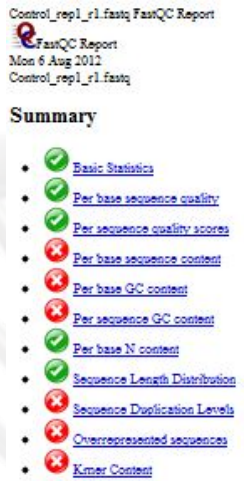




1. Click on "NGS: QC and manipulation"
2. Click on "Fastqc: Fastqc QC
3. Select options:
   a) This is what the window looks like when first opened
   b) Choose fastq file and give it a useful name
4. Click "Execute"
5. Do the exact same thing for the other 3 fastq files

FastQC Output Report

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


Control_rep1_r1.fastq FastQC Report

FastQC Report
Mon 6 Aug 2012
Control_rep1_r1.fastq

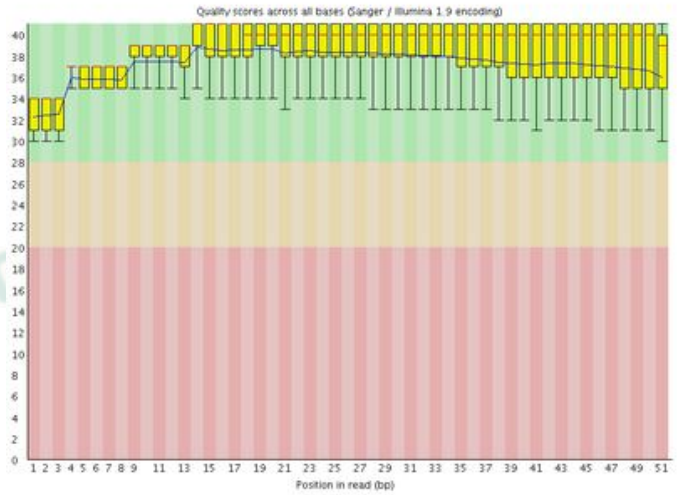
Summary

- 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

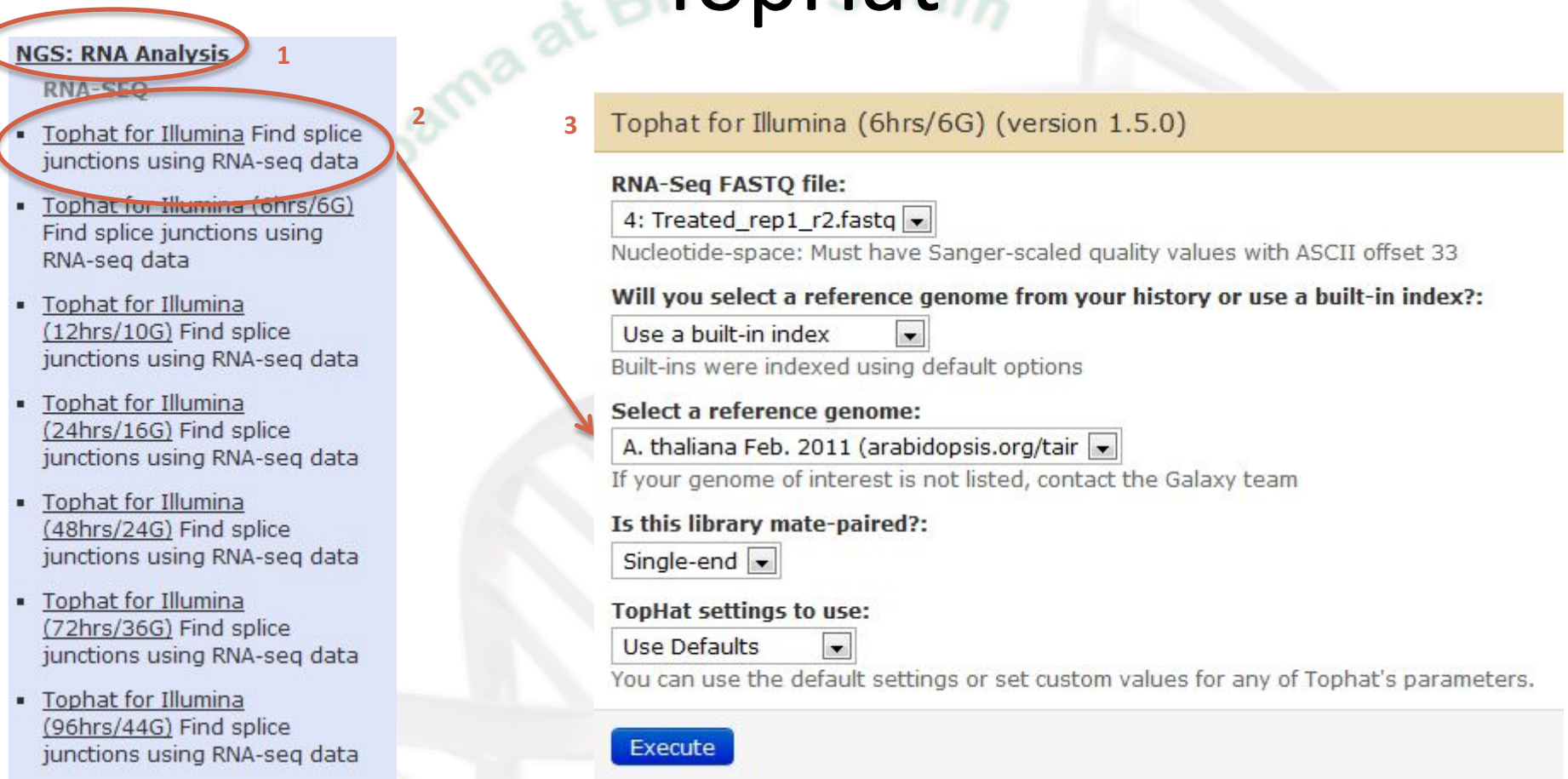
Basic Statistics

Measure	Value
Filename	Control_rep1_r1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	100000
Filtered Sequences	0
Sequence length	51
%GC	46

Per base sequence quality



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 a built-in index"
   b) Select the reference genome
3. Select "Paired-end"
4. Select reverse fastq read file
5. Input "150" (ask sequencing center for this info)
6. Can choose "Commonly used" 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

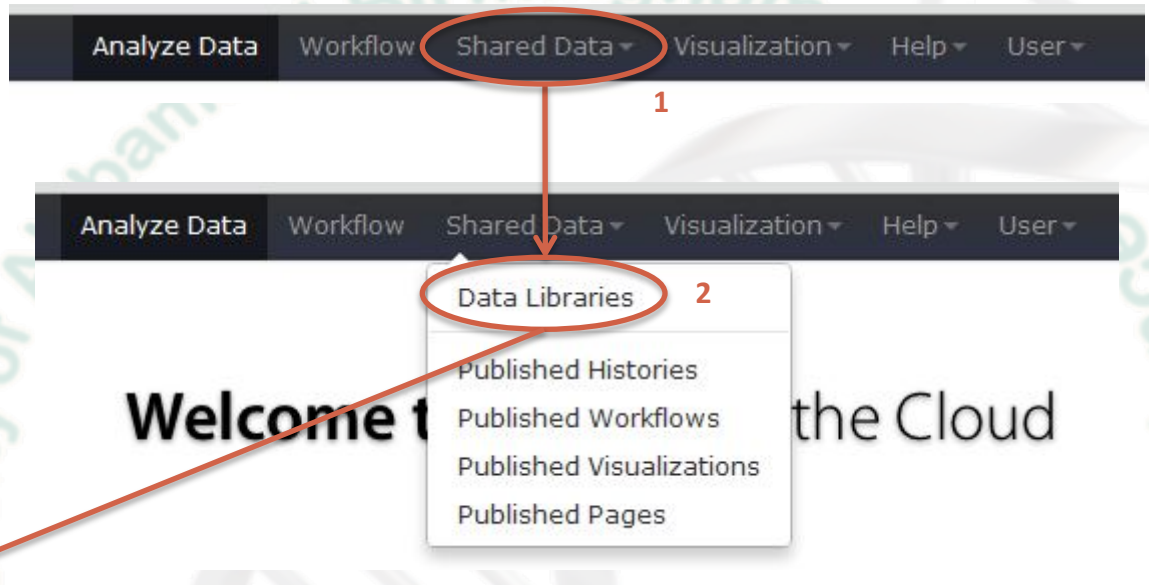
The following job has been successfully added to the queue:

13: Tophat for Illumina (6hrs/6G) on data 2 and data 1: splice junctions

14: Tophat for Illumina (6hrs/6G) on data 2 and data 1: 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.

GTF Annotation Files



3

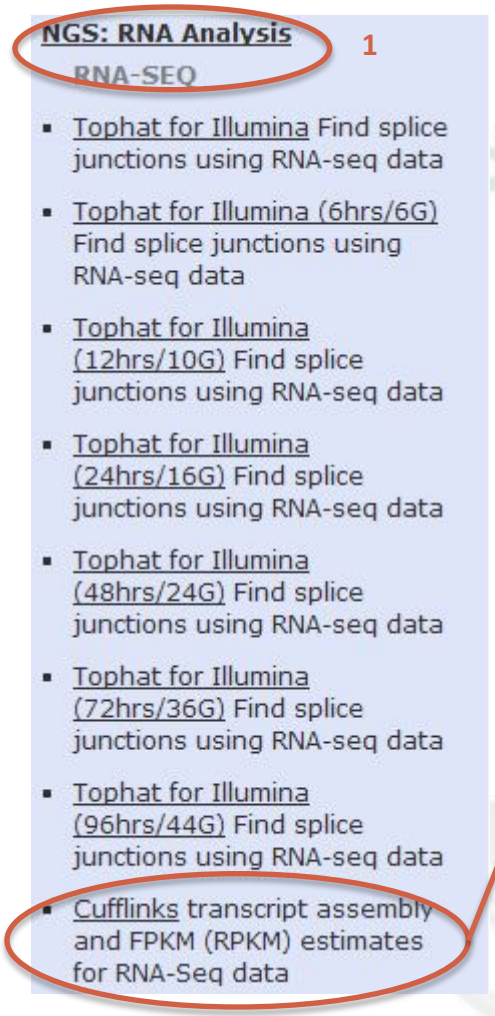
Data Library "Patched GTF annotation files for Cufflinks"

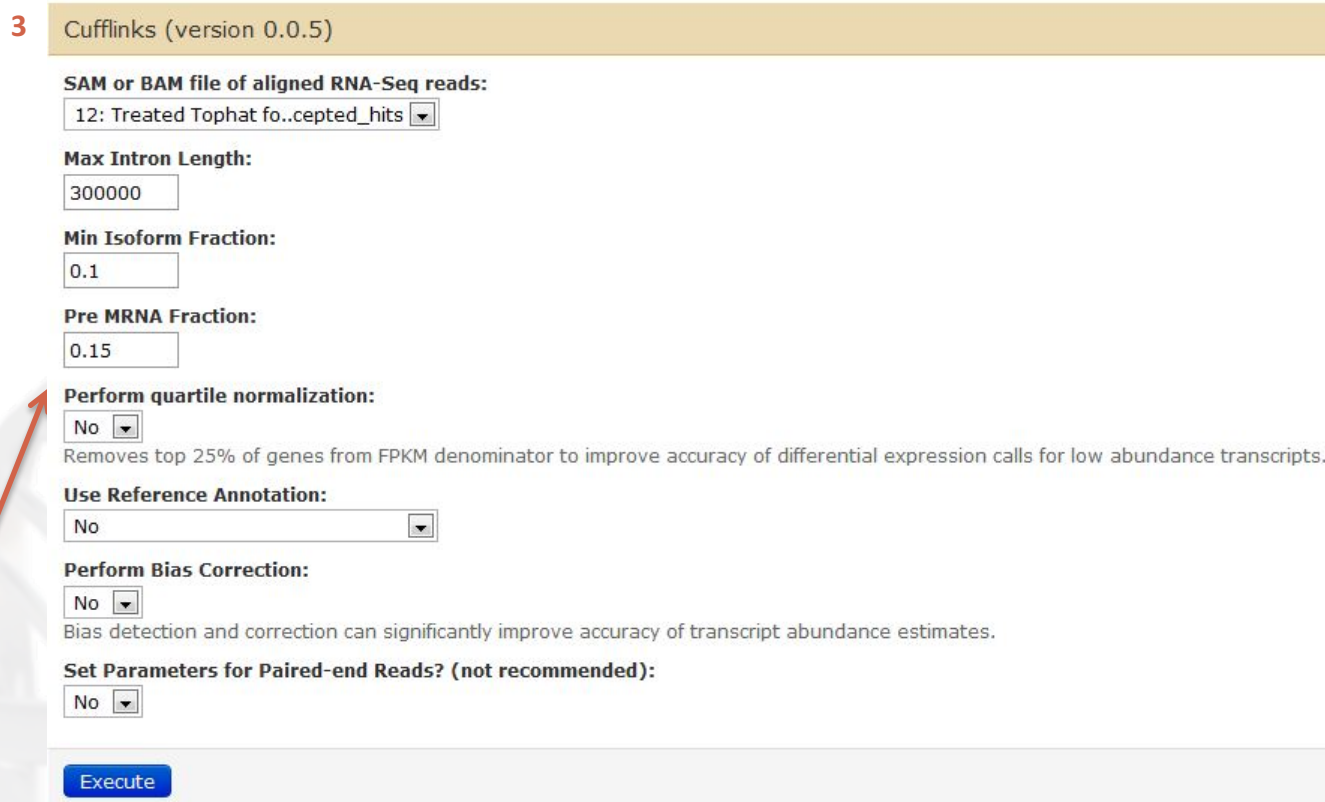
RefGene annotation files patched for Cufflinks in GTF format

Name	Message	Data type	Date uploaded	File size
hg19_RefGene_patched3.gtf	None	gtf	2011-07-22	92.7 Mb
mm9_RefGene_patched3.gtf	None	gtf	2011-07-22	65.5 Mb
rn4_RefGene_patched3.gtf		gtf	2012-02-29	38.4 Mb
Tupaia_belangeri.TREESHREW.63.sorted2.patched.gtf	Not sure if the tupBel1 is the same build as 63!	gtf	2011-08-03	70.4 Mb
Zv9_refGene_patched3.gtf		gtf	2012-02-29	35.6 Mb

For selected datasets: Import to current history Go

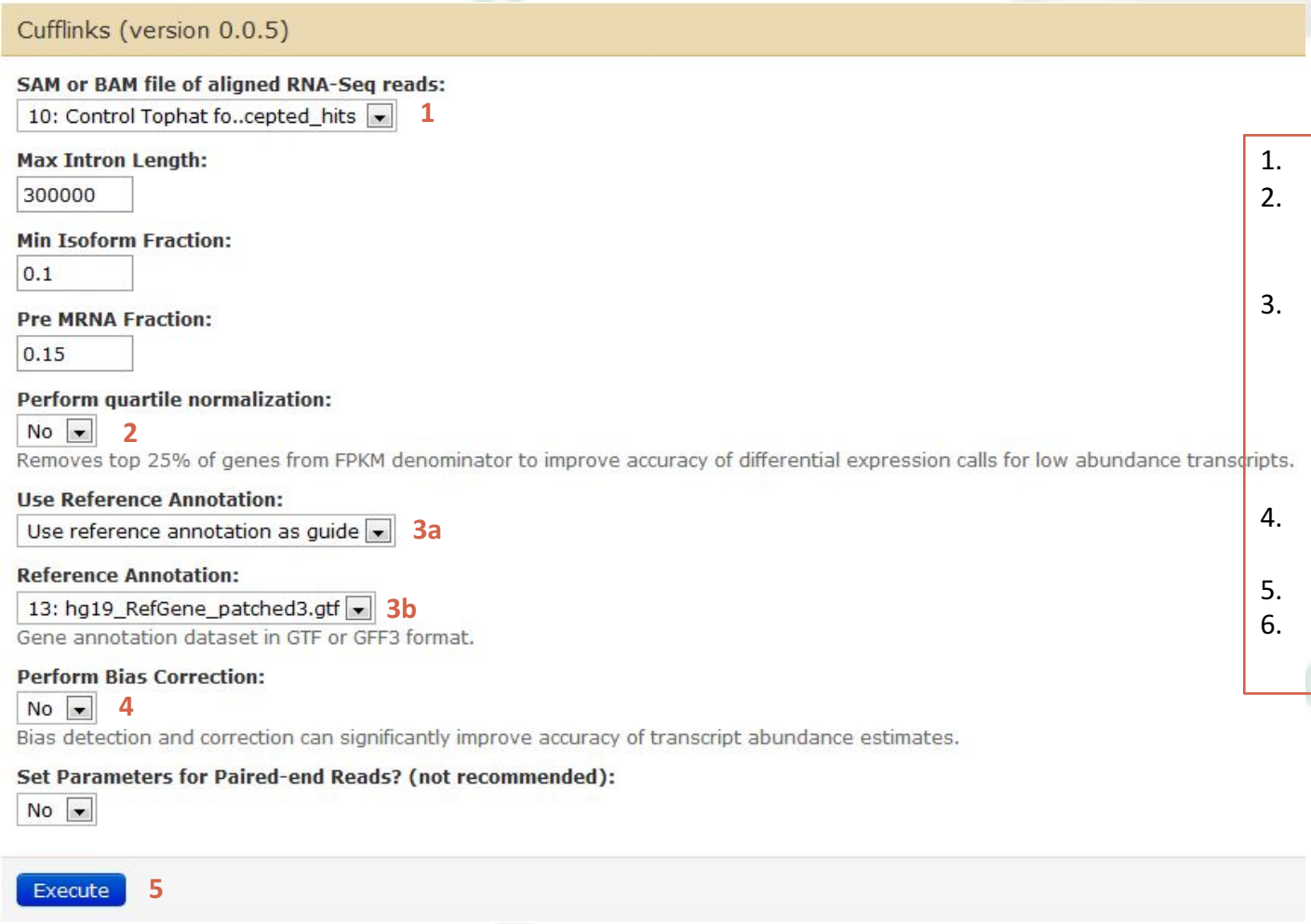
Cufflinks





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 as guide"
   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 Goecks' 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...

Some Cufflinks options to be aware of

-I/–max-intron-length <int>

The maximum intron length. Cufflinks will not report transcripts with introns longer than this, and will ignore SAM alignments with REF_SKIP CIGAR operations longer than this. The default is 300,000.

-F/–min-isoform-fraction <0.0-1.0>

After calculating isoform abundance for a gene, Cufflinks filters out transcripts that it believes are very low abundance, because isoforms expressed at extremely low levels often cannot reliably be assembled, and may even be artifacts of incompletely spliced precursors of processed transcripts. This parameter is also used to filter out introns that have far fewer spliced alignments supporting them. The default is 0.1, or 10% of the most abundant isoform (the major isoform) of the gene.

-j/–pre-mrna-fraction <0.0-1.0>

Some RNA-Seq protocols produce a significant amount of reads that originate from incompletely spliced transcripts, and these reads can confound the assembly of fully spliced mRNAs. Cufflinks uses this parameter to filter out alignments that lie within the intronic intervals implied by the spliced alignments. The minimum depth of coverage in the intronic region covered by the alignment is divided by the number of spliced reads, and if the result is lower than this parameter value, the intronic alignments are ignored. The default is 15%.

Cufflinks output files

The following job has been successfully added to the queue:

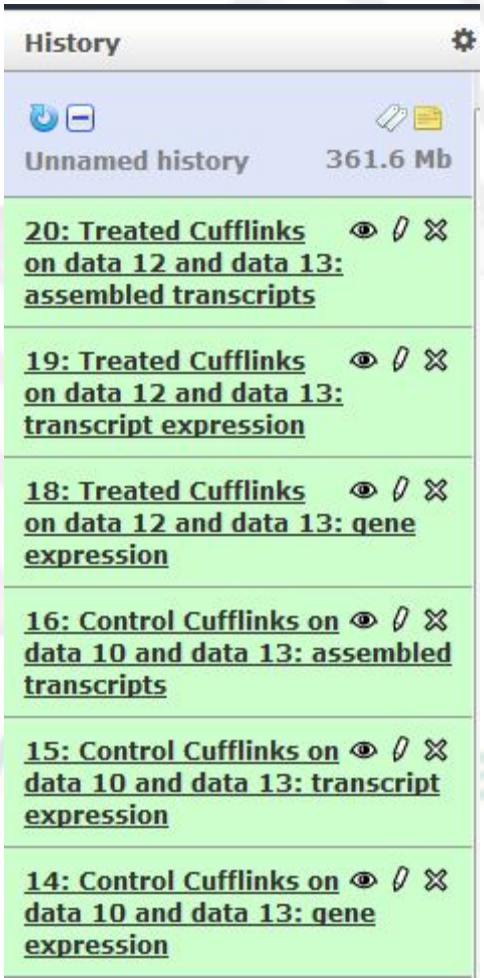
14: Cufflinks on data 10 and data 13: gene expression

15: Cufflinks on data 10 and data 13: transcript expression

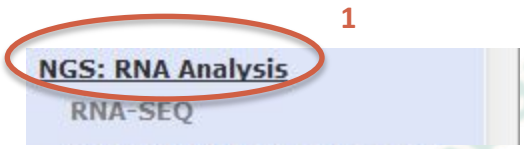
16: Cufflinks on data 10 and data 13: assembled transcripts

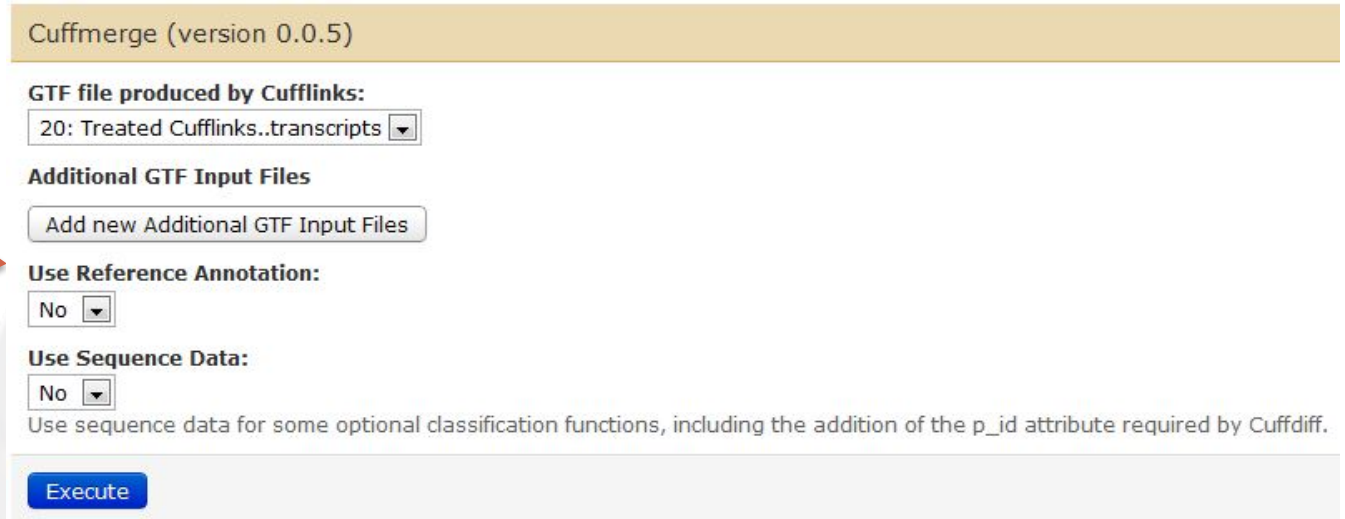
17: Cufflinks on data 10 and data 13: total map mass

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.



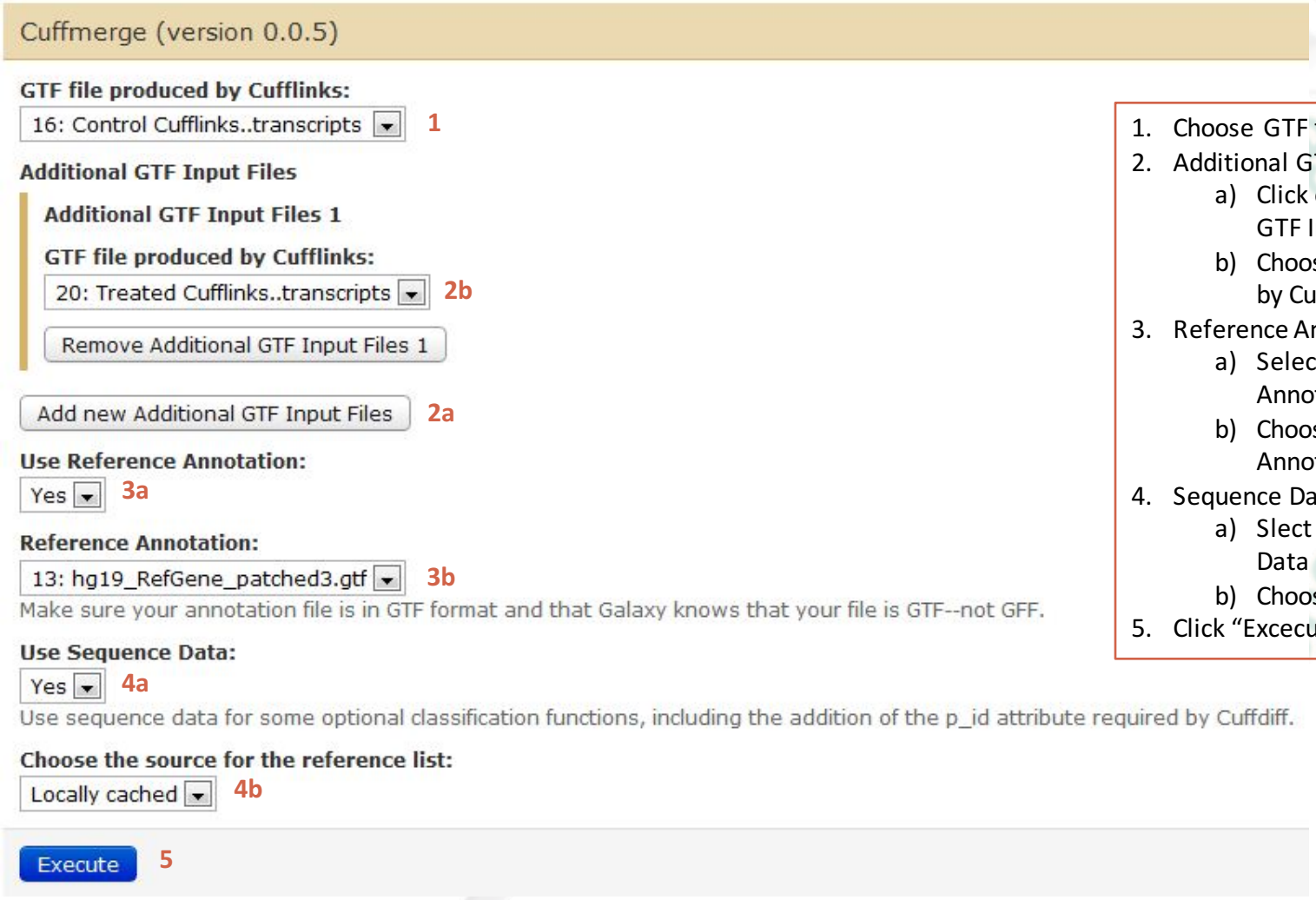
Cuffmerge





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

Cuffmerge



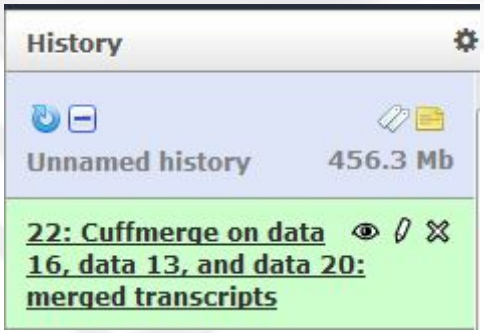
1. Choose GTF file produced by Cufflinks
2. Additional GTF Input Files:
   a) Click on "Add new Additional GTF Input Files"
   b) Choose other GTF file produced by Cufflinks
3. Reference Annotation:
   a) Select "Yes" to Use Reference Annotation
   b) Choose GTF Reference Annotation file from history
4. Sequence Data:
   a) Slect "Yes" to Use Sequence Data
   b) Choose "Locally cached"
5. Click "Excecute"

Cuffmerge output files

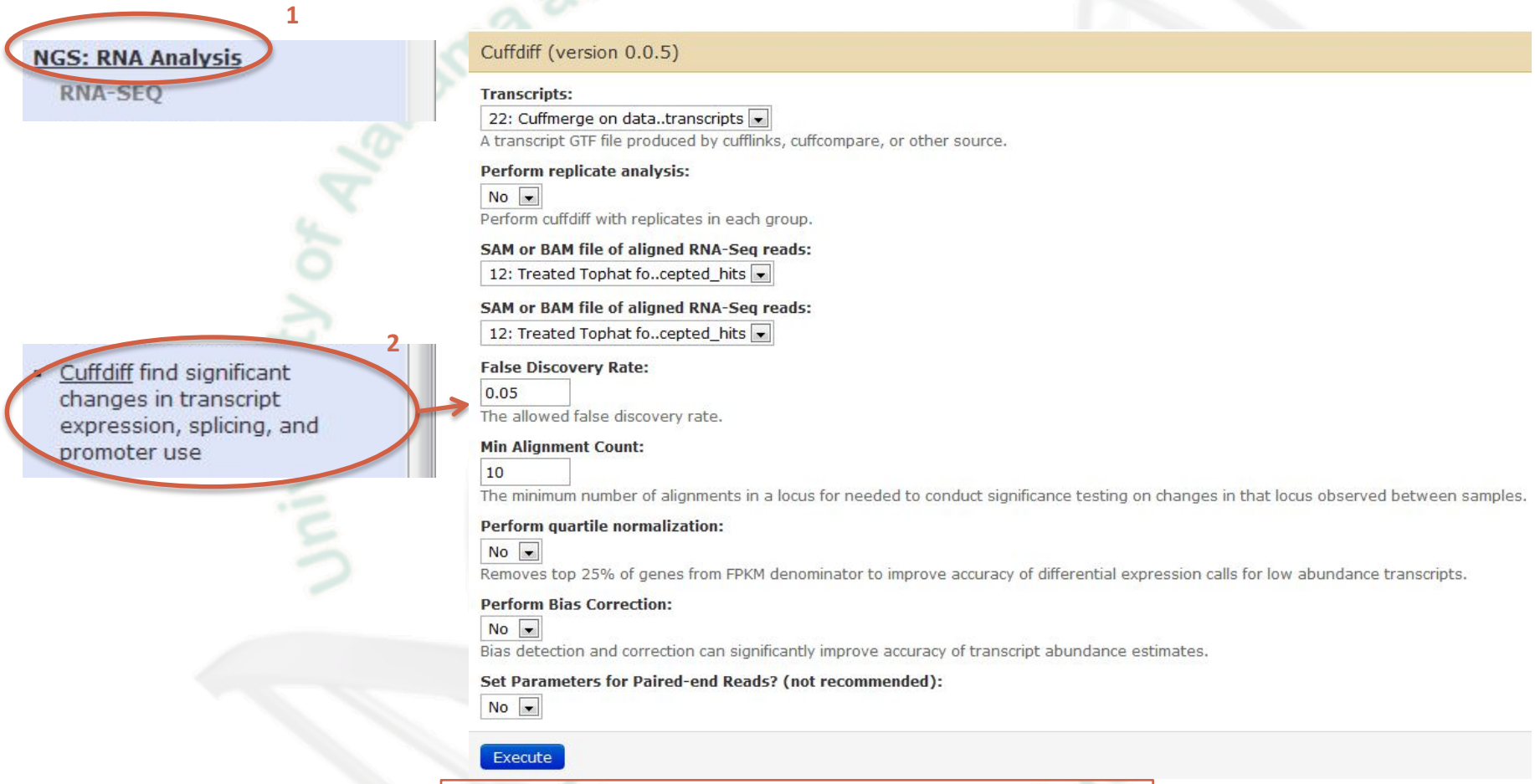
The following job has been successfully added to the queue:

22: Cuffmerge on data 16, data 13, and data 20: 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.

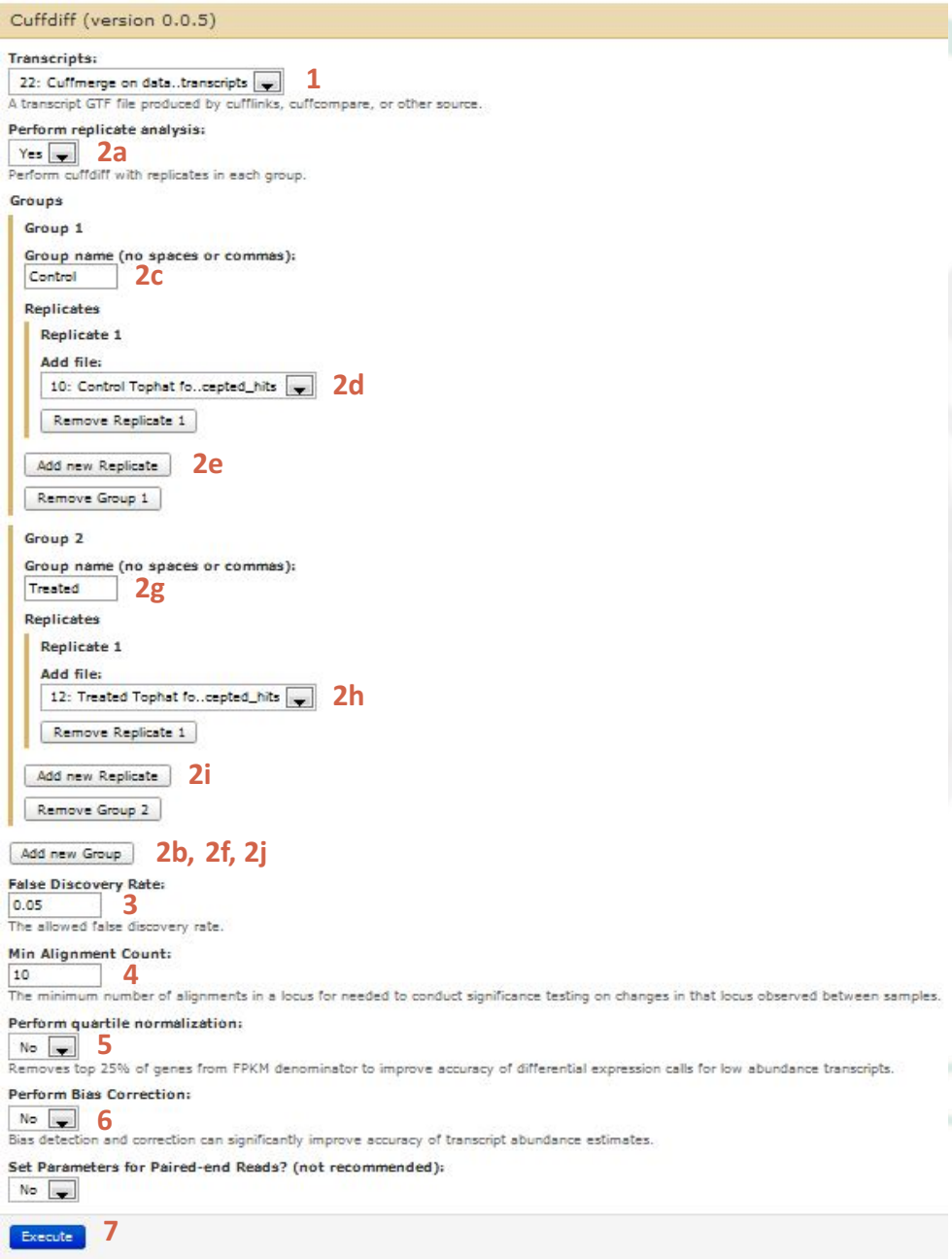


Cuffdiff



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:

23: Cuffdiff on data 12, data 10, and data 22: splicing differential expression testing

24: Cuffdiff on data 12, data 10, and data 22: promoters differential expression testing

25: Cuffdiff on data 12, data 10, and data 22: CDS overloading diffential expression testing

26: Cuffdiff on data 12, data 10, and data 22: CDS FPKM differential expression testing

27: Cuffdiff on data 12, data 10, and data 22: CDS FPKM tracking

28: Cuffdiff on data 12, data 10, and data 22: TSS groups differential expression testing

29: Cuffdiff on data 12, data 10, and data 22: TSS groups FPKM tracking

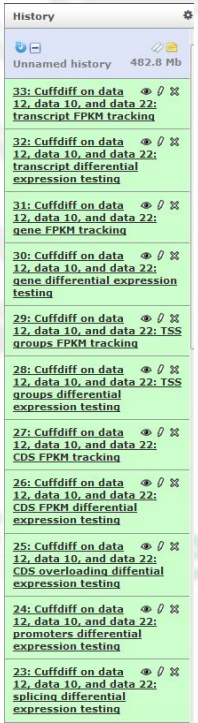
30: Cuffdiff on data 12, data 10, and data 22: gene differential expression testing

31: Cuffdiff on data 12, data 10, and data 22: gene FPKM tracking

32: Cuffdiff on data 12, data 10, and data 22: transcript differential expression testing

33: Cuffdiff on data 12, data 10, and data 22: 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

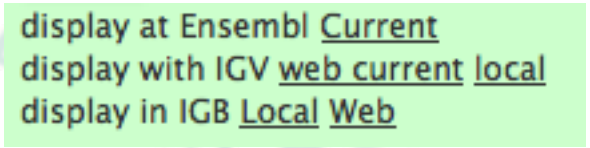
test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
TCONS_00000001	XLOC_000001	OR4F5	chr1:69090-70008	Control	Treated	NOTEST	0	7.91888	1.79769e+308	1.79769e+308	0.369441	1	no
TCONS_00000002	XLOC_000002	LOC100132062	chr1:323891-328581	Control	Treated	OK	6512.86	50.1428	-7.0211	4.36714	1.25886e-05	0.000667762	yes
TCONS_00000003	XLOC_000002	LOC100133331	chr1:323891-328581	Control	Treated	OK	40727.9	1208.59	-5.07462	3.12382	0.00178519	0.0157435	yes
TCONS_00000004	XLOC_000003	OR4F29	chr1:367658-368597	Control	Treated	NOTEST	120.192	11.5757	-3.37617	0.827381	0.408021	1	no
TCONS_00000005	XLOC_000004	LOC643837	chr1:763015-791316	Control	Treated	OK	0	1136.01	1.79769e+308	1.79769e+308	0.0959697	0.130354	no
TCONS_00000006	XLOC_000004	LOC643837	chr1:763015-791316	Control	Treated	LOWDATA	0	0	-1.79769e+308	0	1	1	no
TCONS_00000007	XLOC_000005	SAMD11	chr1:861120-894687	Control	Treated	NOTEST	0	165.375	1.79769e+308	1.79769e+308	0.0784572	1	no
TCONS_00000008	XLOC_000006	KLHL17	chr1:895863-901099	Control	Treated	OK	0	935.161	1.79769e+308	1.79769e+308	0.0958257	0.130354	no
TCONS_00000009	XLOC_000006	KLHL17	chr1:895863-901099	Control	Treated	OK	0	1552.38	1.79769e+308	1.79769e+308	0.098175	0.130354	no
TCONS_00000010	XLOC_000006	KLHL17	chr1:895863-901099	Control	Treated	OK	0	653.036	1.79769e+308	1.79769e+308	0.0842346	0.130354	no
TCONS_00000011	XLOC_000007	PLEKHN1	chr1:901876-917473	Control	Treated	OK	0	259.895	1.79769e+308	1.79769e+308	0.0782193	0.130354	no
TCONS_00000012	XLOC_000007	PLEKHN1	chr1:901876-917473	Control	Treated	NOTEST	0	0	0	0	1	1	no
TCONS_00000013	XLOC_000007	PLEKHN1	chr1:901876-917473	Control	Treated	OK	0	366.221	1.79769e+308	1.79769e+308	0.077757	0.130354	no
TCONS_00000014	XLOC_000007	PLEKHN1	chr1:901876-917473	Control	Treated	NOTEST	0	0	0	0	1	1	no
TCONS_00000015	XLOC_000008	ISG15	chr1:948846-949919	Control	Treated	OK	0	6611.59	1.79769e+308	1.79769e+308	0.0677355	0.130354	no
TCONS_00000016	XLOC_000009	AGRN	chr1:955502-991492	Control	Treated	OK	0	27000.8	1.79769e+308	1.79769e+308	0.215057	0.219233	no
TCONS_00000017	XLOC_000010	LOC254099	chr1:1072396-1079434	Control	Treated	NOTEST	0	0	0	0	1	1	no
TCONS_00000018	XLOC_000011	MIR200B	chr1:1102483-1102578	Control	Treated	NOTEST	0	0	0	0	1	1	no
TCONS_00000019	XLOC_000012	MIR200A	chr1:1103242-1103332	Control	Treated	NOTEST	0	0	0	0	1	1	no
TCONS_00000020	XLOC_000013	MIR429	chr1:1104384-1104467	Control	Treated	NOTEST	0	0	0	0	1	1	no
TCONS_00000021	XLOC_000014	TTLL10	chr1:1109285-1133313	Control	Treated	NOTEST	0	0	0	0	1	1	no
TCONS_00000022	XLOC_000014	TTLL10	chr1:1109285-1133313	Control	Treated	NOTEST	0	0	0	0	1	1	no

Gene differential expression testing output

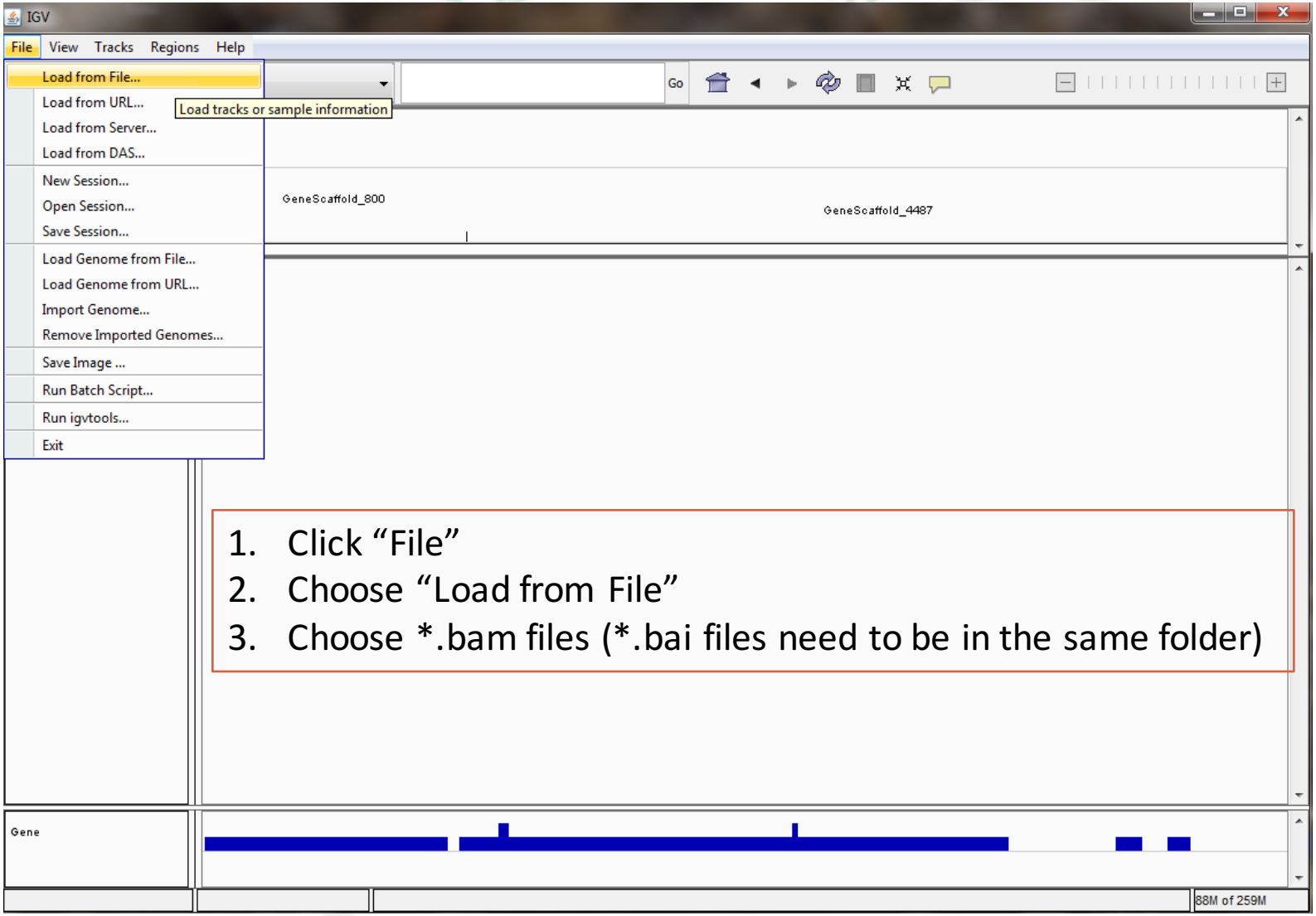
test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
XLOC_000001	XLOC_000001	OR4F5	chr1:69090-70008	Control	Treated	NOTEST	0	7.91888	1.79769e+308	1.79769e+308	0.369441	1	no
XLOC_000002	XLOC_000002	LOC100132062,LOC100133331	chr1:323891-328581	Control	Treated	OK	47240.8	1258.73	-5.22999	3.58623	0.00033549	0.00357856	yes
XLOC_000003	XLOC_000003	OR4F29	chr1:367658-368597	Control	Treated	NOTEST	120.192	11.5757	-3.37617	0.827381	0.408021	1	no
XLOC_000004	XLOC_000004	LOC643837	chr1:763015-791316	Control	Treated	OK	0	1968.53	1.79769e+308	1.79769e+308	0.0161068	0.0355459	yes
XLOC_000005	XLOC_000005	SAMD11	chr1:861120-894687	Control	Treated	NOTEST	0	165.375	1.79769e+308	1.79769e+308	0.0784572	1	no
XLOC_000006	XLOC_000006	KLHL17	chr1:895863-901099	Control	Treated	OK	0	3140.58	1.79769e+308	1.79769e+308	0.00733214	0.0213299	yes
XLOC_000007	XLOC_000007	PLEKHN1	chr1:901876-917473	Control	Treated	OK	0	626.115	1.79769e+308	1.79769e+308	0.0132232	0.0313439	yes
XLOC_000008	XLOC_000008	ISG15	chr1:948846-949919	Control	Treated	OK	0	6611.59	1.79769e+308	1.79769e+308	0.0677355	0.0852164	no
XLOC_000009	XLOC_000009	AGRN	chr1:955502-991492	Control	Treated	OK	0	27000.8	1.79769e+308	1.79769e+308	0.215057	0.218471	no
XLOC_000010	XLOC_000010	LOC254099	chr1:1072396-1079434	Control	Treated	NOTEST	0	0	0	0	1	1	no
XLOC_000011	XLOC_000011	MIR200B	chr1:1102483-1102578	Control	Treated	NOTEST	0	0	0	0	1	1	no
XLOC_000012	XLOC_000012	MIR200A	chr1:1103242-1103332	Control	Treated	NOTEST	0	0	0	0	1	1	no
XLOC_000013	XLOC_000013	MIR429	chr1:1104384-1104467	Control	Treated	NOTEST	0	0	0	0	1	1	no
XLOC_000014	XLOC_000014	TTLL10	chr1:1109285-1133313	Control	Treated	NOTEST	0	0	0	0	1	1	no
XLOC_000015	XLOC_000015	B3GALT6	chr1:1167628-1170420	Control	Treated	OK	0	1211.76	1.79769e+308	1.79769e+308	0.0668946	0.0852164	no
XLOC_000016	XLOC_000016	SCNN1D	chr1:1215815-1227409	Control	Treated	NOTEST	0	74.5236	1.79769e+308	1.79769e+308	0.0721728	1	no
XLOC_000017	XLOC_000017	PUSL1	chr1:1243993-1260046	Control	Treated	OK	0	2317.82	1.79769e+308	1.79769e+308	0.0649866	0.0852164	no
XLOC_000018	XLOC_000018	GLTPD1	chr1:1260142-1264276	Control	Treated	OK	0	1597.74	1.79769e+308	1.79769e+308	0.0669804	0.0852164	no
XLOC_000019	XLOC_000019	TAS1R3	chr1:1266725-1269844	Control	Treated	NOTEST	0	31.2299	1.79769e+308	1.79769e+308	0.0912112	1	no
XLOC_000020	XLOC_000020	LOC148413	chr1:1334909-1342693	Control	Treated	OK	0	2591.73	1.79769e+308	1.79769e+308	0.101067	0.109708	no
XLOC_000021	XLOC_000021	TMEM88B	chr1:1361507-1363167	Control	Treated	NOTEST	0	0	0	0	1	1	no
XLOC_000022	XLOC_000022	VWA1	chr1:1370902-1378262	Control	Treated	NOTEST	0	4.59925	1.79769e+308	1.79769e+308	0.230105	1	no
XLOC_000023	XLOC_000023	ATAD3C	chr1:1385068-1405538	Control	Treated	OK	0	270.979	1.79769e+308	1.79769e+308	0.0615518	0.0852164	no
XLOC_000024	XLOC_000024	ATAD3B	chr1:1407163-1431582	Control	Treated	OK	0	9725.9	1.79769e+308	1.79769e+308	0.0932631	0.106586	no
XLOC_000025	XLOC_000025	ATAD3A	chr1:1447522-1470067	Control	Treated	OK	0	15128.3	1.79769e+308	1.79769e+308	0.125562	0.131737	no
XLOC_000026	XLOC_000026	MIB2	chr1:1550794-1565990	Control	Treated	OK	0	1139.11	1.79769e+308	1.79769e+308	0.00159396	0.00822516	yes

Using IGV to view the data

- http://www.broadinstitute.org/igv/
- Several ways to view the accepted_hits.bam file from TopHat:
  - Download the bam file to your computer (don't forget to download the bam_index file (*.bai) and then load into IGV
  - View them directly from Galaxy (no downloading required)



Load aligned BAM files into IGV

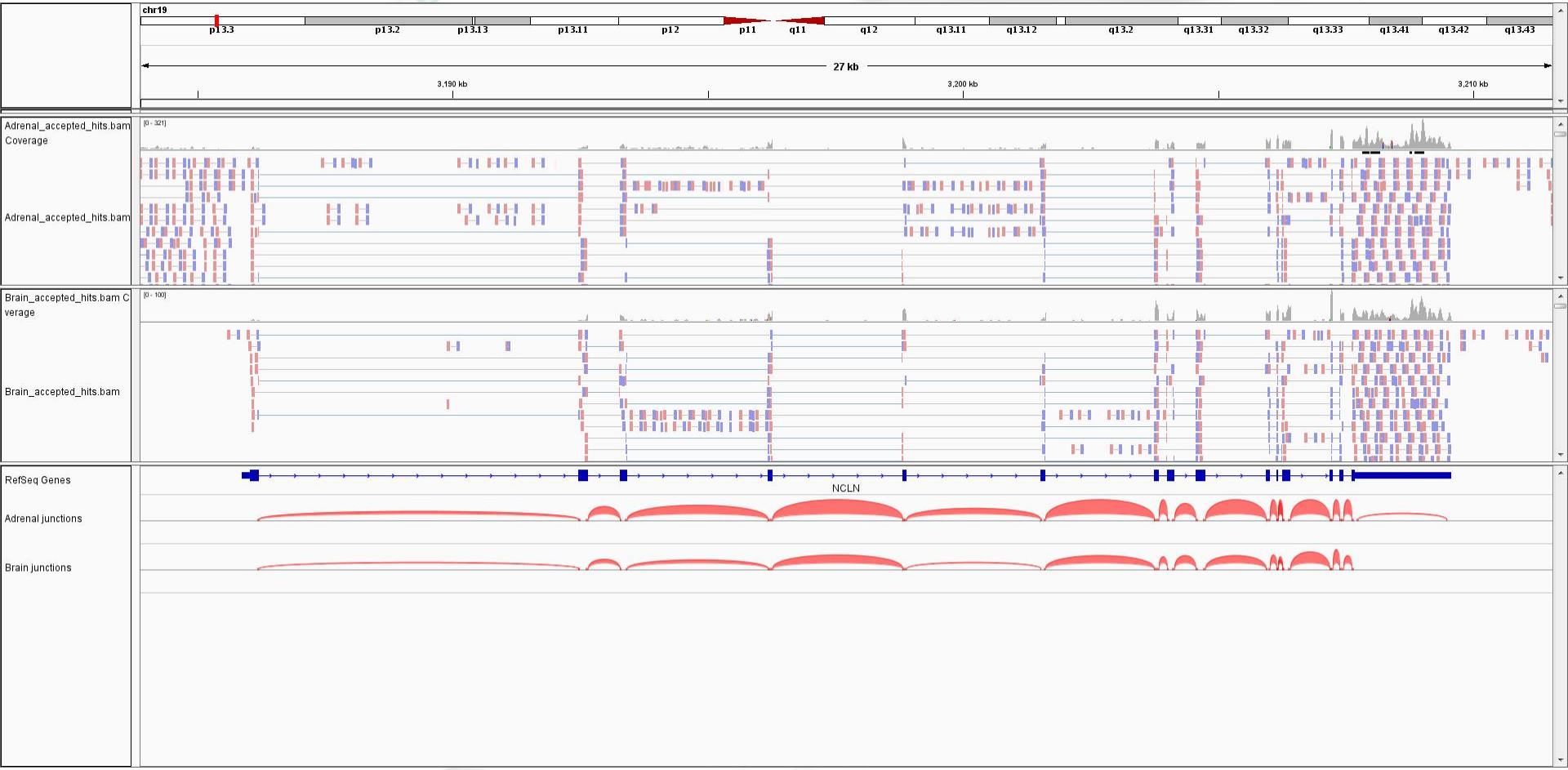


1. Click “File”
2. Choose “Load from File”
3. Choose *.bam files (*.bai files need to be in the same folder)

chr19


chr19
p13.3
p13.2
p13.13
p13.11
p12
p11
q11
q12
q13.11
q13.12
q13.2
q13.31
q13.32
q13.33
q13.41
q13.42
q13.43
572 kb
3,000 kb
3,100 kb
3,200 kb
3,300 kb
3,400 kb
3,500 kb
Adrenal_accepted_hits.bam Coverage
Adrenal_accepted_hits.bam
Zoom in to see alignments.
Brain_accepted_hits.bam Coverage
Brain_accepted_hits.bam
Zoom in to see alignments.
RefSeq Genes
TLE6
TLE2
AES
GNA11
GNA15
S1PR4
NCLN
CELF5
NFIC
C19orf77
FZR1
MFSD12
Adrenal junctions
Brain junctions

NCLN


chr19
p13.3
p13.2
p13.13
p13.11
p12
p11
q11
q12
q13.11
q13.12
q13.2
q13.31
q13.32
q13.33
q13.41
q13.42
q13.43
27 kb
3,190 kb
3,200 kb
3,210 kb
Adrenal_accepted_hits.bam Coverage
[0 - 321]
Adrenal_accepted_hits.bam
Brain_accepted_hits.bam Coverage
[0 - 100]
Brain_accepted_hits.bam
RefSeq Genes
NCLN
Adrenal junctions
Brain junctions

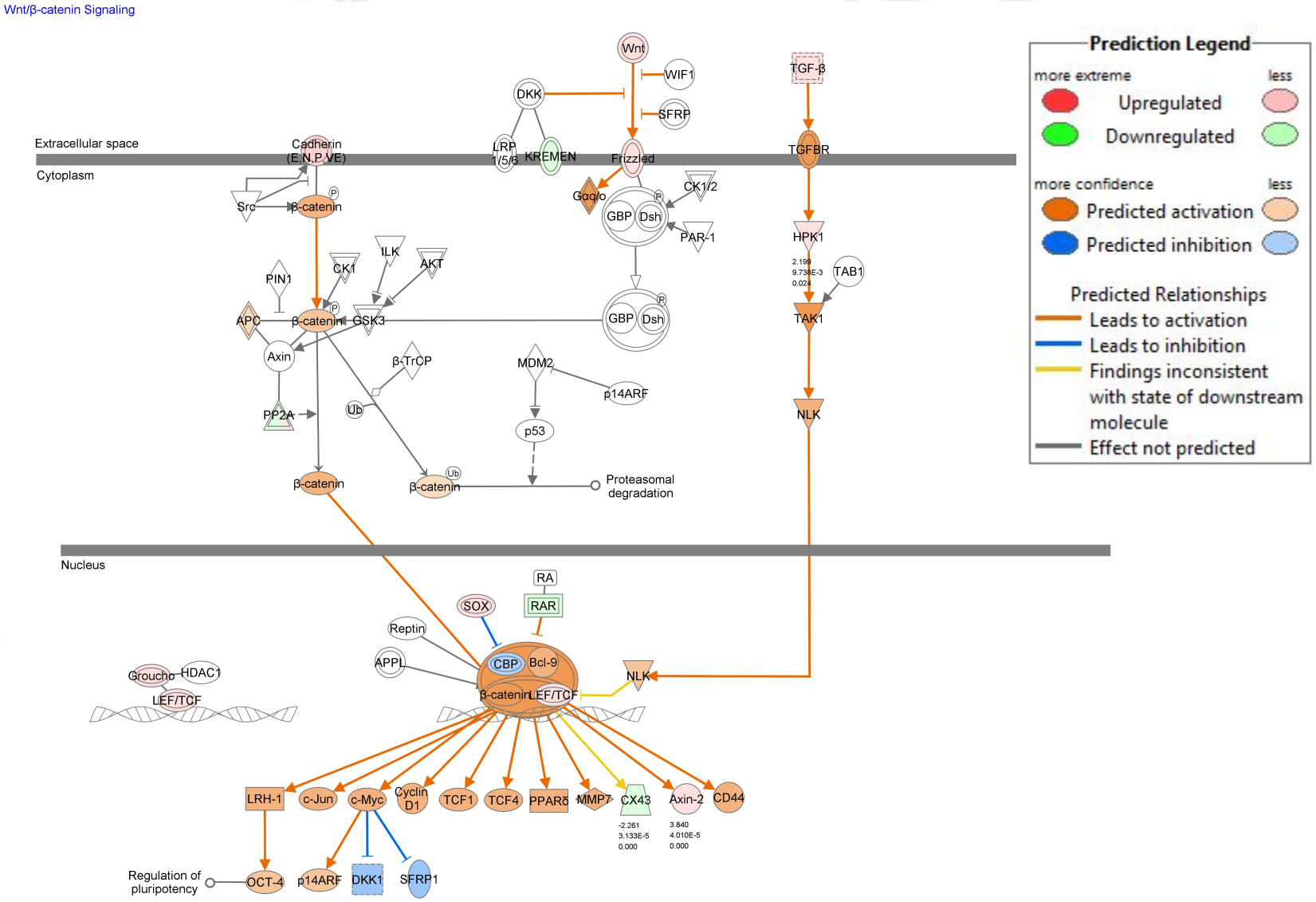
What to do with your list of genes

- Apply a Systems Biology approach to data mine and analyze your data
- Tools and databases available (some free, others $$) to define the underlying biology behind different –omics data
- These tools and databases will identify and prioritize the most relevant pathways, networks and cellular processes affected by your dataset.

Pathways & Ontology Analysis Tools

Tool	Link	Price
Reactome	http://www.reactome.org	Free
IPA	http://ingenuity.com/	$$$
GeneGo Metacore	http://www.genego.com/	$$$
Cytoscape	http://www.cytoscape.org/	Free
GenMAPP	http://www.genmapp.org/	Free
InterMine	http://intermine.org/	Free
KEGG	http://www.genome.jp/kegg/	Free
GO	http://www.geneontology.org/	Free
Panther	http://www.pantherdb.org/	Free
DAVID	http://david.abcc.ncifcrf.gov/	Free
And many, many more!!		

Network Example



Variant Pathway Tools

Tool	Link	Price
Ingenuity Variant Analysis	http://ingenuity.com/	$$$
Clinical Genomics Toolkit	http://lsresearch.thomsonreuters.com/pages/solutions/21/clinical-genomics-toolkit	$$$
Cartagenia	http://www.cartagenia.com/	$$$
Tute Genomics	http://tutegenomics.com/	$$$
VariantStudio	http://www.illumina.com/informatics/research/biological-data-interpretation/variantstudio.html	$$$

Demo

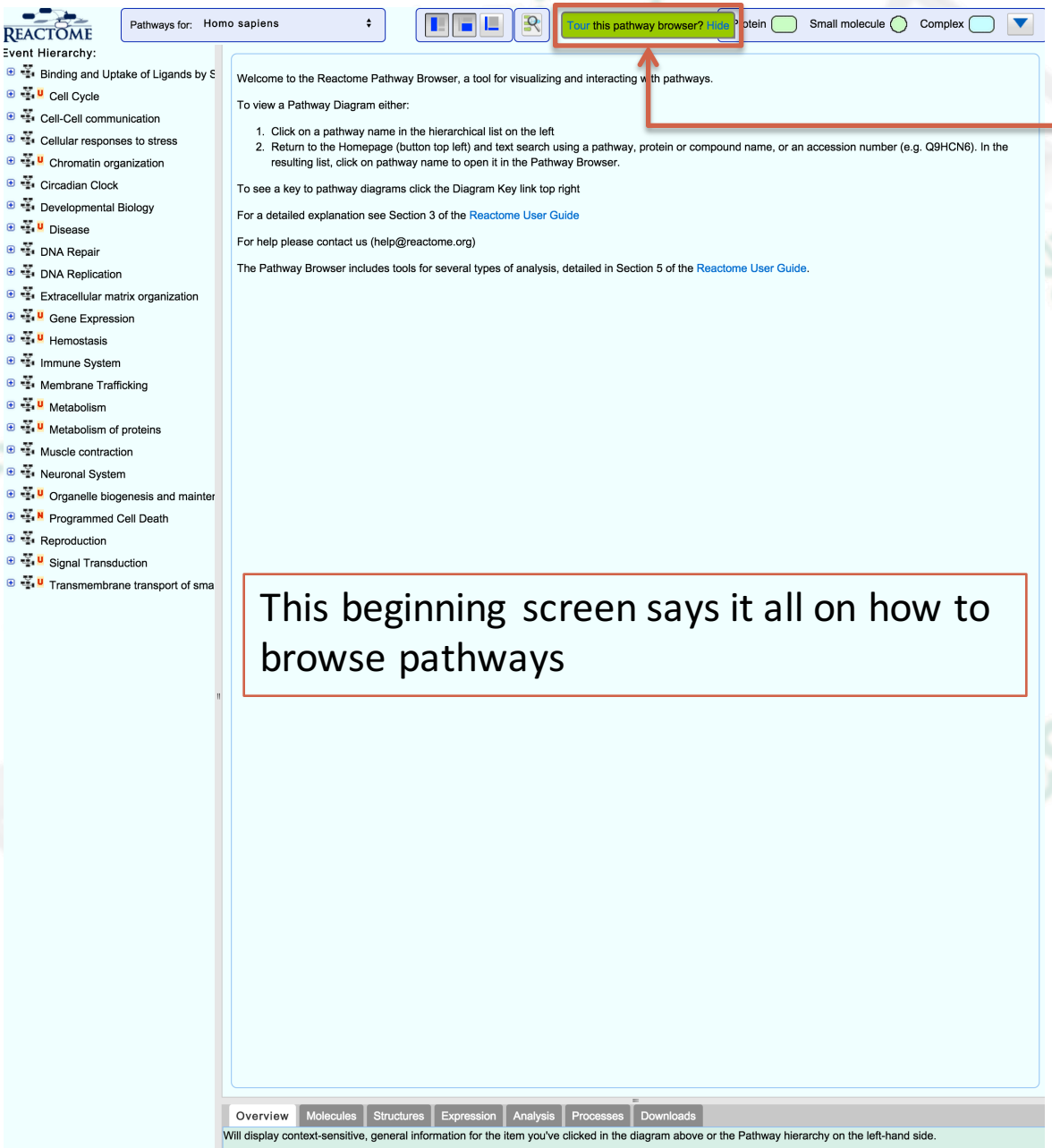
REACTOME

Reactome

- Open-source, open access.
- Manually curated.
- Peer-reviewed pathway database (pathway annotations are authored by “expert” biologists).
- Some of the tools they have:
  - Browse pathways
  - Map IDs to pathways
  - Overrepresentation analysis
  - Compare species
  - Analyze expression data


Quick links to tools most commonly used.
REACTOME
A CURATED PATHWAY DATABASE
About
Content
Documentation
Tools
Community
Download
Contact
e.g. O95631, NTN1, signalir
Search
Manual & tutorials
Other useful tools
Browse Pathways
Analyze Data
New
Reactome FI Network
User Guide
Data Download
Contact Us
Tweets
Current Version: Reactome V51
Tweets by @reactome
About Reactome
Reactome is a free, open-source, curated and peer reviewed pathway database. Our goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education. The current version (v51) of Reactome was released on December 8, 2014.
OICR
Ontario Institute for Cancer Research
NYU Langone MEDICAL CENTER
CSH
Cold Spring Harbor Laboratory
EMBL-EBI
The development of Reactome is supported by a grant from the US National Institutes of Health (P41 HG003751), Ontario Research Fund, and the European Molecular Biology Laboratory.
About
About Reactome
News
Reactome Team
Scientific Advisory Board
Other Reactomes
License Agreement
Reactome Disclaimer
Content
Table of Contents
DOIs
Data Schema
Editorial Calendar
Statistics
ORCID Integration Project
Documentation
User Guide
Data Model
Orthology Prediction
Object/Relational Mapping
Wiki
Linking to Reactome
Referencing Reactome
Tools
Pathway Browser
Analyze Data
Species Comparison
Reactome FI Network
Advanced Search
Author/Reviewer Search
Small Molecule Search
Analysis Service
Community
Reactome Outreach
Reactome Events
Reactome Training
Reactome Publications
Papers Citing Reactome
Resources Guide
Mailing List
Download
Contact

Browse Pathways



Helpful tutorial

This beginning screen says it all on how to browse pathways

Analyze Data

Analysis Tools

This tool merges pathway identifier mapping, overrepresentation and expression analysis into a single tabbed data analysis portal, with integrated visualization and summary features.

Select a file from your computer and click on the "Analyse" button to perform the analysis.

Select data file for analysis 1b [Choose File] No file chosen ☑ Project to human 2 [Analyse]

▼ Click here to paste your data or try example data sets...

Paste the data to analyse

1a

1. Upload data to Reactome:
   a. Paste, or
   b. Choose file from computer (tab-delimited), or
   c. For demo purposes, click “Example”
2. Click “Analyse”

Some examples: 1c

- Uniprot accession list
- Gene name list
- Gene NCBI / Entrez list
- Small molecules (ChEBI)
- Small molecules (KEGG)
- Microarray data
- Metabolomics data

[Clear] ☑ Project to human 2 [Analyse]

Analyze Data Results

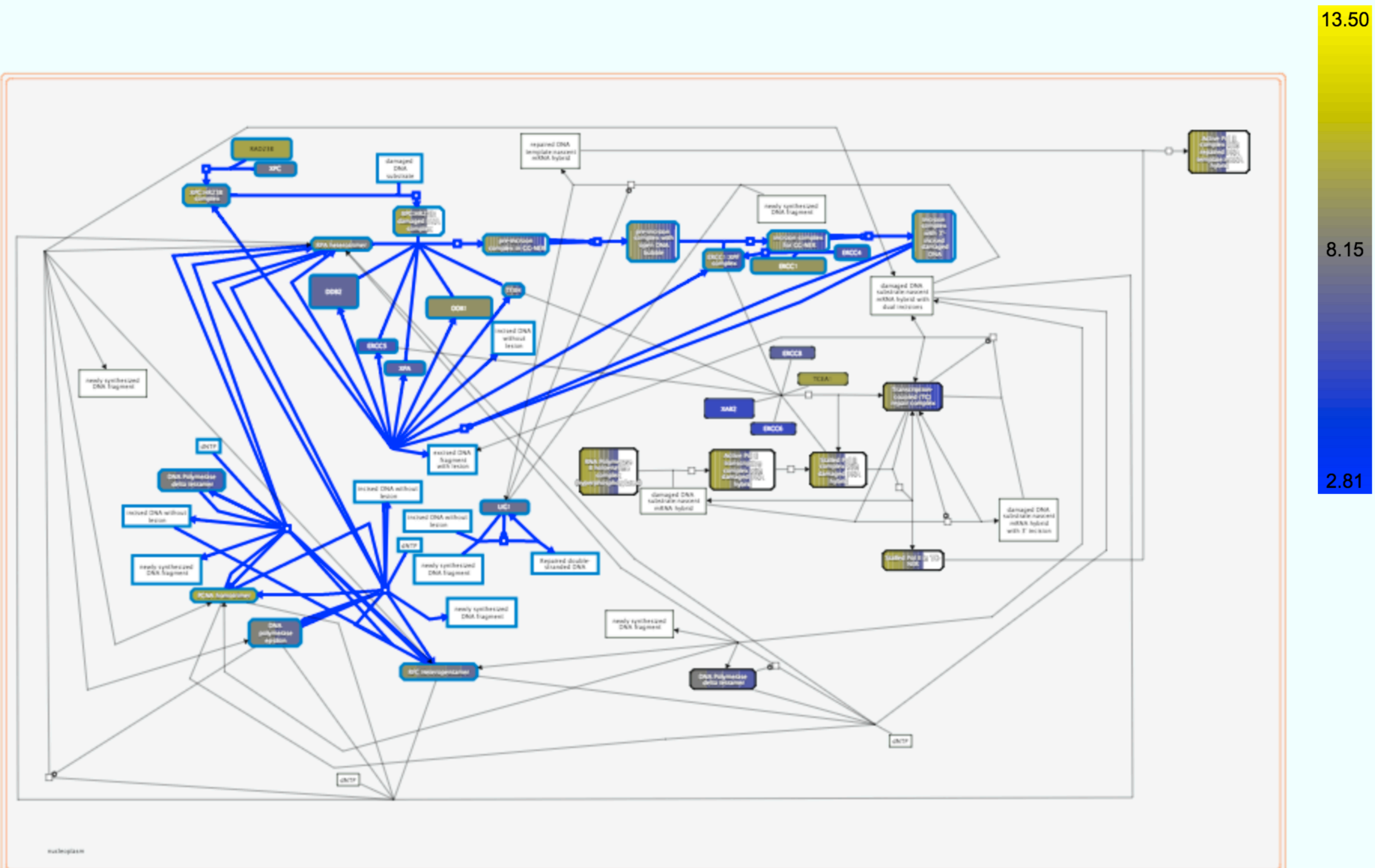
Overview | Molecules | Structures | Expression | Analysis (973) | Processes | Downloads

Results for: UNIPROT (973) | Type: Expression | [Data: Probeset] | Results | Identifiers not found: 29

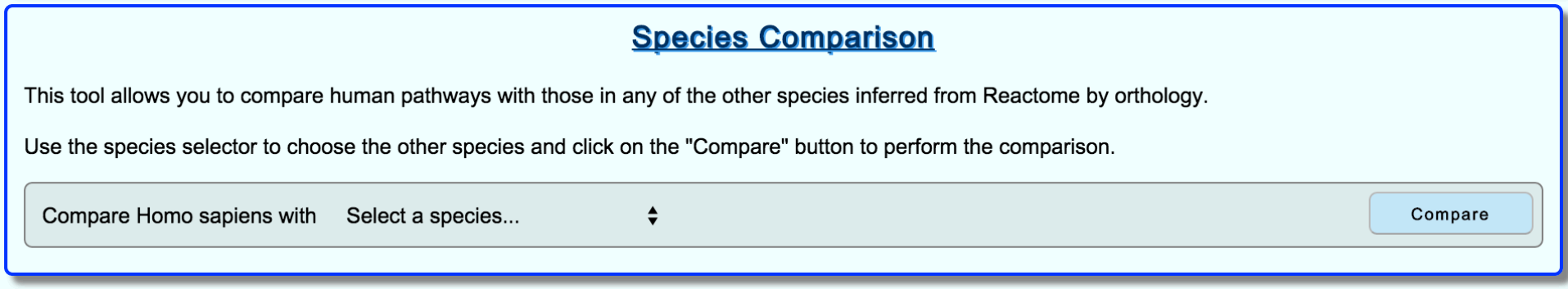
Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total
Global Genomic NER (GG-NER)	34	34	0.004	1.11E-16	1.33E-15	11	11
Nucleotide Excision Repair	46	50	0.006	1.11E-16	1.33E-15	21	21
Transcription-coupled NER (TC-NER)	41	45	0.006	1.11E-16	1.33E-15	10	10
RNA Polymerase II Transcription Elongation	37	42	0.005	1.11E-16	1.33E-15	8	8
Formation of HIV elongation complex in the absence of HIV Tat	37	42	0.005	1.11E-16	1.33E-15	2	2
Formation of RNA Pol II elongation complex	37	42	0.005	1.11E-16	1.33E-15	2	2
Formation of the HIV-1 Early Elongation Complex	28	32	0.004	1.11E-16	1.33E-15	5	5
Formation of the Early Elongation Complex	28	32	0.004	1.11E-16	1.33E-15	3	3
Activation of the pre-replicative complex	26	30	0.004	1.11E-16	1.33E-15	8	8
HIV Transcription Elongation	36	42	0.005	1.11E-16	1.33E-15	15	15
Tat-mediated elongation of the HIV-1 transcript	36	42	0.005	1.11E-16	1.33E-15	8	8
Formation of HIV-1 elongation complex containing HIV-1 Tat	36	42	0.005	1.11E-16	1.33E-15	5	5
RNA Polymerase II Pre-transcription Events	53	62	0.008	1.11E-16	1.33E-15	17	17
Transcription of the HIV genome	55	65	0.008	1.11E-16	1.33E-15	45	45
RNA Polymerase II Transcription Initiation And Promoter Clearance	35	43	0.005	1.11E-16	1.33E-15	9	9
RNA Polymerase II HIV Promoter Escape	35	43	0.005	1.11E-16	1.33E-15	7	7
RNA Polymerase II Transcription Pre-Initiation And Promoter Opening	35	43	0.005	1.11E-16	1.33E-15	5	5
RNA Polymerase II Promoter Escape	35	43	0.005	1.11E-16	1.33E-15	5	5
HIV Transcription Initiation	35	43	0.005	1.11E-16	1.33E-15	4	4
RNA Polymerase II Transcription Initiation	35	43	0.005	1.11E-16	1.33E-15	3	3

Table can be downloaded in various formats

Clicking any of the pathway names will show the respective pathway


13.50
8.15
2.81
RAD23B
XPC
damaged DNA substrate
repaired DNA template:nascent mRNA hybrid
newly synthesized DNA fragment
RPA heterotrimer
DDB2
DDB1
ERCC5
XPA
incised DNA without lesion
excised DNA fragment with lesion
ERCC4
ERCC1
damaged DNA template:nascent mRNA hybrid with dual incisions
ERCC8
TCEA1
XAB2
ERCC6
Transcription-coupled (TC) repair complex
damaged DNA template:nascent mRNA hybrid
damaged DNA template:nascent mRNA hybrid with 3' incision
dNTP
DNA Polymerase delta tetramer
PCNA homotrimer
DNA polymerase epsilon
RFC heteropentamer
LIG1
Repaired double stranded DNA
nucleoplasm

Compare species



This tool will allow you to compare human pathways to any other species pathways they have in their database

Compare species

Overview | Molecules | Structures | Expression | Analysis (1,610) | Processes | Downloads

Results for: TOTAL (1610) | Type: Species Comparison | Results

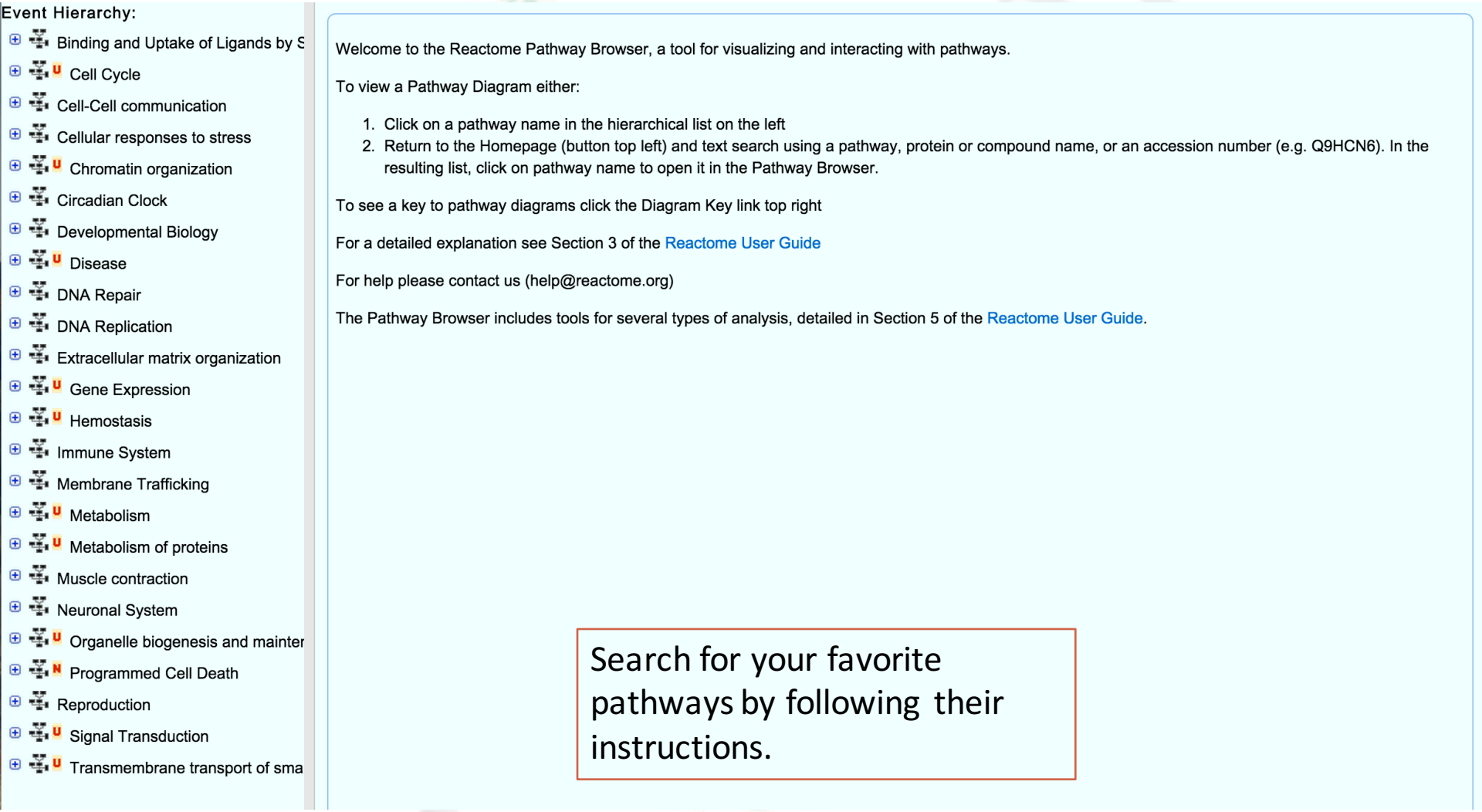
Pathway name	Entities found	Entities Total	Entities ratio	Entities pValue	Entities FDR	Reactions found	Reactions total
Signal Transduction	1827	2,372	0.245	3.06E-12	1.16E-8	*1,388*	*1,408*
Hemostasis	474	561	0.058	2.06E-7	2.8E-4	*299*	*305*
Cell Cycle, Mitotic	425	496	0.051	2.21E-7	2.8E-4	*272*	*272*
Signalling by NGF	288	325	0.034	1.72E-6	1.63E-3	*170*	*170*
Separation of Sister Chromatids	171	179	0.018	4.66E-6	3.54E-3	*8*	*8*
Cell Cycle	470	574	0.059	5.69E-6	3.6E-3	*345*	*345*
Mitotic Metaphase and Anaphase	180	194	0.02	1.35E-5	6.87E-3	*12*	*12*
Mitotic Anaphase	179	193	0.02	1.45E-5	6.87E-3	*11*	*11*
Platelet activation, signaling and aggregation	216	248	0.026	8.47E-5	3.05E-2	*106*	*109*
Signaling by Rho GTPases	120	125	0.013	9.21E-5	3.05E-2	*5*	*5*
Rho GTPase cycle	120	125	0.013	9.21E-5	3.05E-2	*5*	*5*
Assembly of the primary cilium	173	193	0.02	1.08E-4	3.05E-2	*50*	*50*
NGF signalling via TRKA from the plasma membrane	208	239	0.025	1.18E-4	3.05E-2	*116*	*116*
Cell Cycle Checkpoints	116	121	0.012	1.25E-4	3.05E-2	*38*	*38*
M Phase	258	306	0.032	1.37E-4	3.05E-2	*63*	*63*
Peptide ligand-binding receptors	173	194	0.02	1.38E-4	3.05E-2	*66*	*68*
Axon guidance	298	360	0.037	1.45E-4	3.05E-2	*239*	*239*
Downstream signaling events of B Cell Receptor (BCR)	173	195	0.02	1.75E-4	3.49E-2	*37*	*37*
Mitotic G1-G1/S phases	129	139	0.014	2.03E-4	3.83E-2	*53*	*53*
mRNA Splicing	112	119	0.012	3.12E-4	5.36E-2	*14*	*14*

Result | Mapping

1-20 of 1,610

Here, I chose mouse and as before, clicking on the pathway name will open the pathway map, and the table/results can be downloaded.

Browse pathways



Search for your favorite pathways by following their instructions.

Demo

PANTHER

Panther

http://pantherdb.org/

- Tools and data on the PANTHER site can be used to:
  - Get information about a gene of interest
  - Explore protein families, molecular functions, biological processes, cellular components and pathways
  - Generate lists of genes that belong to a given protein family or subfamily, have a given molecular function or participate in a given biological process or pathway, e.g. generate a candidate gene list for a disease
  - Analyze lists of genes, proteins or transcripts according to categories based on family, molecular function, biological process, cellular component or pathway, e.g. analyze mRNA microarray data

Now includes comprehensive GO annotations directly imported from the GO database

Search

All

Go

Quick links

Whole genome function views

Genome statistics

How to cite PANTHER

NEW! Recent publication describing PANTHER

News

PANTHER gene analysis tools now support comprehensive GO annotations.

Click for additional info.

Newsletter subscription

Enter your Email:

Subscribe

Gene List Analysis | Browse | Sequence Search | cSNP Scoring | Keyword Search

Please refer to our article in Nature Protocols for detailed instructions on how to use this page.

Help Tips

Steps:

1. Select list and list type to analyze
2. Select Organism
3. Select operation

1. **Enter ids and or select file for batch upload. Else enter ids or select file or list from workspace for comparing to a reference list.**

Enter IDs: Supported IDs

separate IDs by a space or comma

Upload IDs: File format

Choose File No file chosen

Please login to be able to select lists from your workspace.

Select List Type:

- (•) ID List
- ( ) Previously exported text search results
- ( ) Workspace list
- ( ) PANTHER Generic Mapping File

2. **Select organism.**

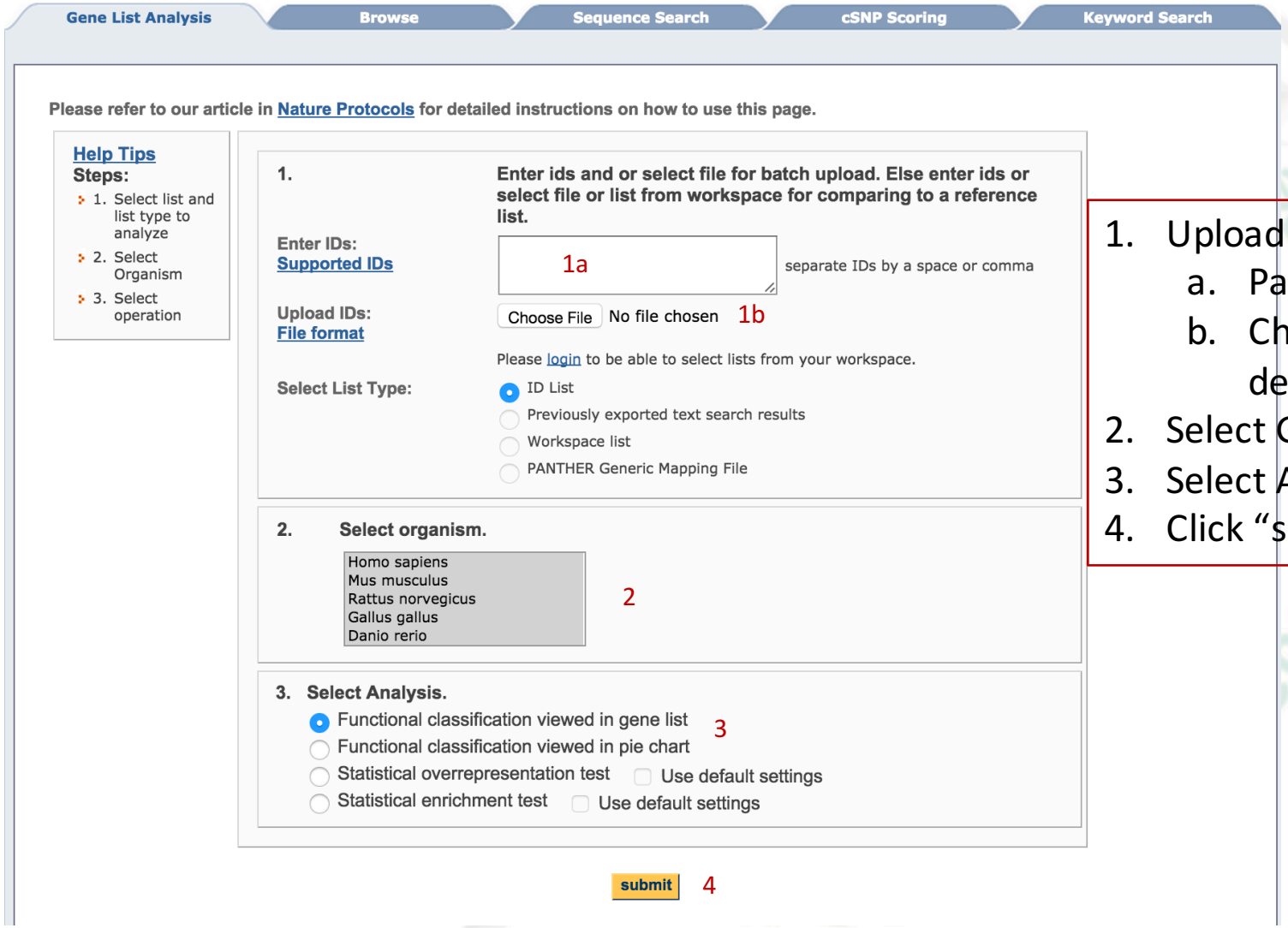
- Homo sapiens
- Mus musculus
- Rattus norvegicus
- Gallus gallus
- Danio rerio

3. **Select Analysis.**

- (•) Functional classification viewed in gene list
- ( ) Functional classification viewed in pie chart
- ( ) Statistical overrepresentation test [ ] Use default settings
- ( ) Statistical enrichment test [ ] Use default settings

submit

Upload multiple gene IDs



1. Upload data to Panther:
   a. Paste your IDs, or
   b. Choose a file (tab-delimited)
2. Select Genome(s)
3. Select Analysis type
4. Click “submit”

- Table listing all the genes in your dataset (scroll to the right to see all the columns
- Table can be downloaded by the "Send list to:" dropdown box
- Click the pie chart icon to view the GO categories in pie charts

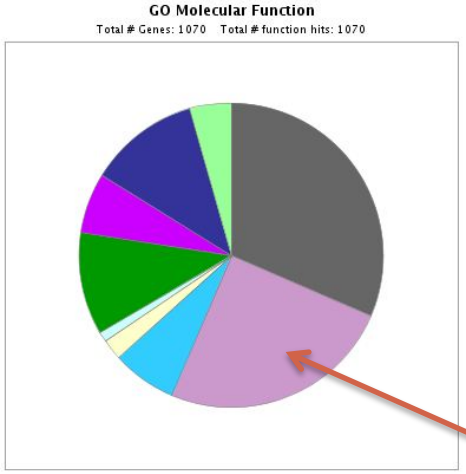
PANTHER GENE LIST Add/remove columns from Gene list view

Convert List to: -Select- Send list to: -Select-

Display: 30 items per page Refine Search

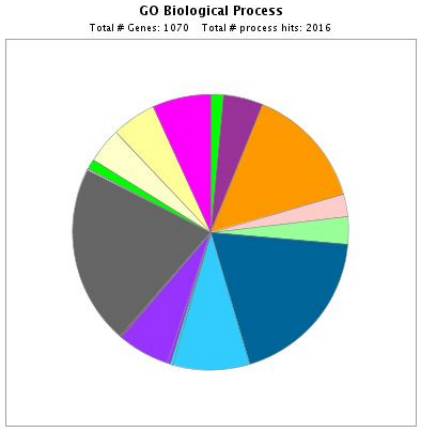
Hits 1-30 of 186 [page: (1) 2 3 4 5 6 7] Number of mapped ids found 164 IDs not found (9)

clr all	Gene ID	Mapped IDs	Gene Name Gene Symbol Ortholog	PANTHER Family/Subfamily	PANTHER Protein Class	Species
1.	HUMAN|Ensembl=ENSG00000134321|UniProtKB=Q8WXG1	RSAD2	Radical S-adenosyl methionine domain-containing protein 2 RSAD2 ortholog	RADICAL S-ADENOSYL METHIONINE DOMAIN-CONTAINING PROTEIN 2 (PTHR21339:SF0)	-	Homo sapiens
2.	HUMAN|Ensembl=ENSG00000205413|UniProtKB=Q5K651	SAMD9	Sterile alpha motif domain-containing protein 9 SAMD9 ortholog	STERILE ALPHA MOTIF DOMAIN-CONTAINING PROTEIN 9 (PTHR16155:SF17)	-	Homo sapiens
3.	HUMAN|Ensembl=ENSG00000121594|UniProtKB=P33681	CD80	T-lymphocyte activation antigen CD80 CD80 ortholog	T-LYMPHOCYTE ACTIVATION ANTIGEN CD80 (PTHR25462:SF214)	ubiquitin-protein ligase	Homo sapiens
4.	HUMAN|Ensembl=ENSG00000241534|UniProtKB=P00751	CFB	Complement factor B CFB ortholog	No Panther Hit	-	Homo sapiens
5.	HUMAN|Ensembl=ENSG00000121858|UniProtKB=P50591	TNFSF10	Tumor necrosis factor ligand superfamily member 10 TNFSF10 ortholog	TUMOR NECROSIS FACTOR LIGAND SUPERFAMILY MEMBER 10 (PTHR11471:SF27)	tumor necrosis factor family member	Homo sapiens
6.	HUMAN|Ensembl=ENSG00000158517|UniProtKB=P14598	NCF1	Neutrophil cytosol factor 1 NCF1 ortholog	No Panther Hit	-	Homo sapiens
7.	HUMAN|Ensembl=ENSG00000189060|UniProtKB=P07305	H1F0	Histone H1.0 H1F0	HISTONE H1.0 (PTHR11467:SF10)	histone	Homo sapiens

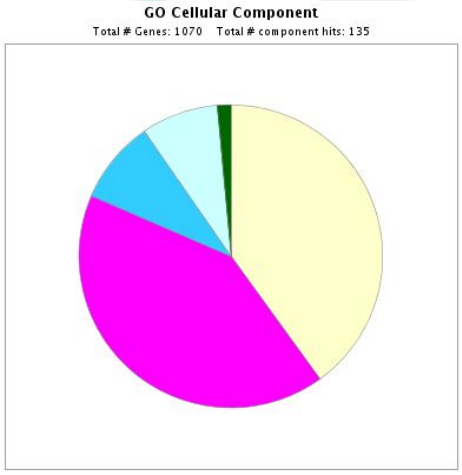


Click to get gene list for a category:
- binding (GO:0005488)
- catalytic activity (GO:0003824)
- enzyme regulator activity (GO:0030234)
- ion channel activity (GO:0005216)
- motor activity (GO:0003774)
- receptor activity (GO:0004872)
- structural molecule activity (GO:0005198)
- transcription regulator activity (GO:0030528)
- transporter activity (GO:0005215)

Color picker powered by Web Colors by VisiBone



Click to get gene list for a category:
- apoptosis (GO:0006915)
- cell adhesion (GO:0007155)
- cell communication (GO:0007154)
- cell cycle (GO:0007049)
- cellular component organization (GO:0016043)
- cellular process (GO:0009987)
- developmental process (GO:0032502)
- generation of precursor metabolites and energy (GO:0006091)
- homeostatic process (GO:0042592)
- immune system process (GO:0002376)
- localization (GO:0051179)
- metabolic process (GO:0008152)
- regulation of biological process (GO:0050789)
- reproduction (GO:0000003)
- response to stimulus (GO:0050896)
- system process (GO:000300
- transport (GO:0006810)



Click to get gene list for a category:
- extracellular region (GO:0005576)
- intracellular (GO:0005622)
- plasma membrane (GO:0005886)
- protein complex (GO:0043234)
- ribonucleoprotein complex (GO:0030529)

Color picker powered by Web Colors by VisiBone

- Pie charts of the 3 GO categories:
  - Molecular Function
  - Biological Process
  - Cellular Component
- Each “wedge” can be clicked on to drill further down into the cateogry
- Clicking on any category name link will list a table of molecules from your dataset found in that particular category


University of Alabama at Birmingham
Heflin Center for Genomic Science

Demo

DAVID

DAVID

http://david.abcc.ncifcrf.gov/

- **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**)
- Provides a comprehensive set of functional annotation tools to better understand the biological meaning behind large datasets.
- Tools:
  - GO
  - Pathways
  - Gene-disease associations
  - Protein functional domains and motifs
  - And much, much more!
- Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature Protocols 4, 44-57 (2009).

DAVID Bioinformatics Resources 6.7

National Institute of Allergy and Infectious Diseases (NIAID), NIH

Shortcut to DAVID Tools

Functional Annotation

Gene-annotation enrichment analysis, functional annotation clustering , BioCarta & KEGG pathway mapping, gene-disease association, homologue match, ID translation, literature match and more

Gene Functional Classification

Provide a rapid means to reduce large lists of genes into functionally related groups of genes to help unravel the biological content captured by high throughput technologies. More

Gene ID Conversion

Convert list of gene ID/accessions to others of your choice with the most comprehensive gene ID mapping repository. The ambiguous accessions in the list can also be determined semi-automatically. More

Gene Name Batch Viewer

Display gene names for a given gene list; Search functionally related genes within your list or not in your list; Deep links to enriched detailed information. More

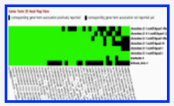
Recommending: A paper published in ***Nature Protocols*** describes step-by-step procedure to use DAVID!

Welcome to DAVID 6.7

2003 - 2014

The **D**atabase for **A**nnotation, **V**isualization and **I**ntegrated **D**iscovery (**DAVID**) v6.7 is an update to the sixth version of our original web-accessible programs. DAVID now provides a comprehensive set of functional annotation tools for investigators to understand biological meaning behind large list of genes. For any given gene list, DAVID tools are able to:

- Identify enriched biological themes, particularly GO terms
- Discover enriched functional-related gene groups
- Cluster redundant annotation terms
- Visualize genes on BioCarta & KEGG pathway maps
- Display related many-genes-to-many-terms on 2-D view.
- Search for other functionally related genes not in the list
- List interacting proteins
- Explore gene names in batch
- Link gene-disease associations
- Highlight protein functional domains and motifs
- Redirect to related literatures
- Convert gene identifiers from one type to another.
- And more


Screen Shot 1


Screen Shot 2

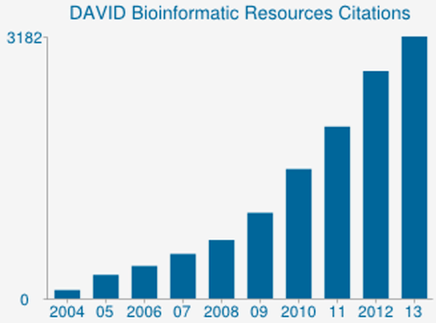

Screen Shot 3

Search

What's Important in DAVID?

- Current (v 6.7) release note
- New requirement to cite DAVID
- IDs of Affy Exon and Gene arrays supported
- Novel Classification Algorithms
- Pre-built Affymetrix and Illumina backgrounds
- User's customized gene background
- Enhanced calculating speed

Statistics of DAVID



- > 10,000 Citations
- Daily Usage: ~1200 gene lists/sublists from ~400 unique researchers.
- Total Usage: ~800,000 gene lists/sublists from >5,000 research institutes world-wide

Please cite *Nature Protocols 2009; 4(1):44 & Nucleic Acids Res. 2009;37(1):1* within any publication that makes use of any methods inspired by **DAVID**.

References and web links

- Galaxy
  - Public website: https://main.g2.bx.psu.edu/
  - UAB: https://www.uab.edu/galaxy
- 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