Transcriptome Analysis

NGS FASTQ file format

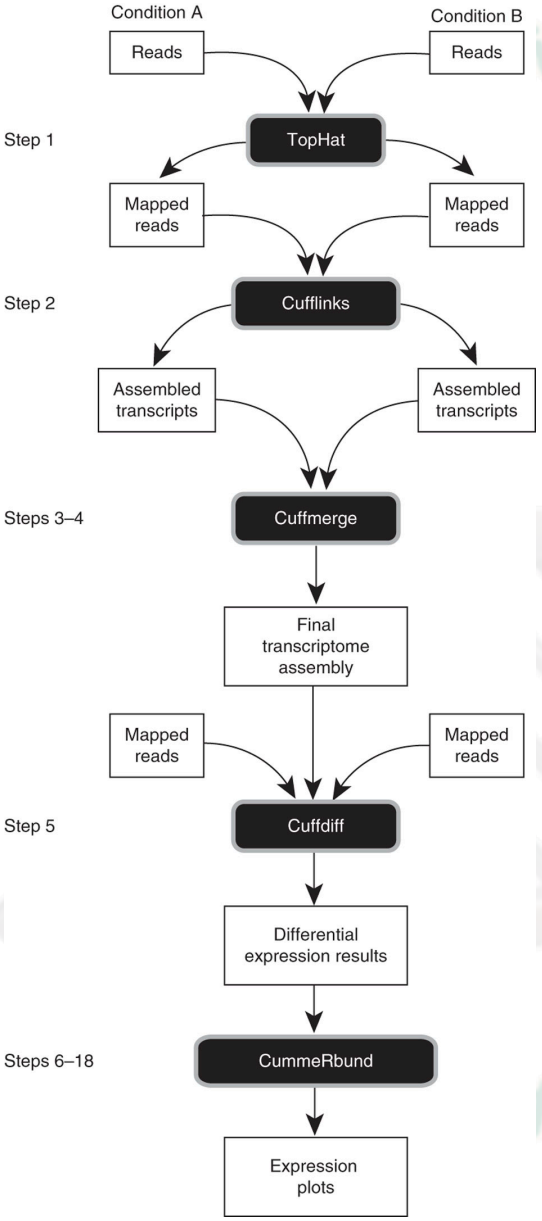
```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....................
 ............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....................
 ..............................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
  |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

Line1: Begins with '@' and followed by a sequence identifier and optional description
Line2: Raw sequence letters
Line3: '+'
Line4: Encodes the quality values for the sequence in Line2 (see above figure)
Repeat Lines1-4 format again and again and again…

```
1 @D5VG2KN1:116:C0NTMACXX:5:1101:1606:2077 2:N:0:GTGAAA
2 CTTNNCTTCATGTNCCTTTCCTCTCATGTCTTCCCTGAGGTCCTCGTAATC
3 +
4 B@@##2=2AFDHH#2<CDHHGIII9HHIIEFF:CEHB@DGHGIIIDGEIEH
  @D5VG2KN1:116:C0NTMACXX:5:1101:1584:2079 2:N:0:GTGAAA
  GGGNNTTCATGATNAAGATGAGAGTGCACGGCTTCTCCTCTGAGAAGGACT
  +
  @?;##22=AD84D#2<<;CDH@HG<C>FHGDBFGEH??DBFGEBB<9CEFC
  @D5VG2KN1:116:C0NTMACXX:5:1101:1526:2088 2:N:0:GTGAAA
  TTTNGCAGCACGGCTTTGTCCTCTGGGGTGAGGGCTGGTGTGGGTAGGGCA
  +
  BBB#4=DDBHHHFIJIJIJJGHEGGIJJIJIJJJGIJJIJHIHJGGJGHFE
  @D5VG2KN1:116:C0NTMACXX:5:1101:1730:2093 2:N:0:GTGAAA
  CCCCCAGGCCAGGTAGCCCAAGCCAAGTGTCCAGAGGTTGACCCTGTGCGT
  +
  CC@FFFFFHHHHHIJJIIJIJIJJJJJGIIGIJJHJJHHJIJJJJJJJJIG
```

RNA-Seq pipeline



Bowtie
Extremely fast, general purpose short read aligner

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

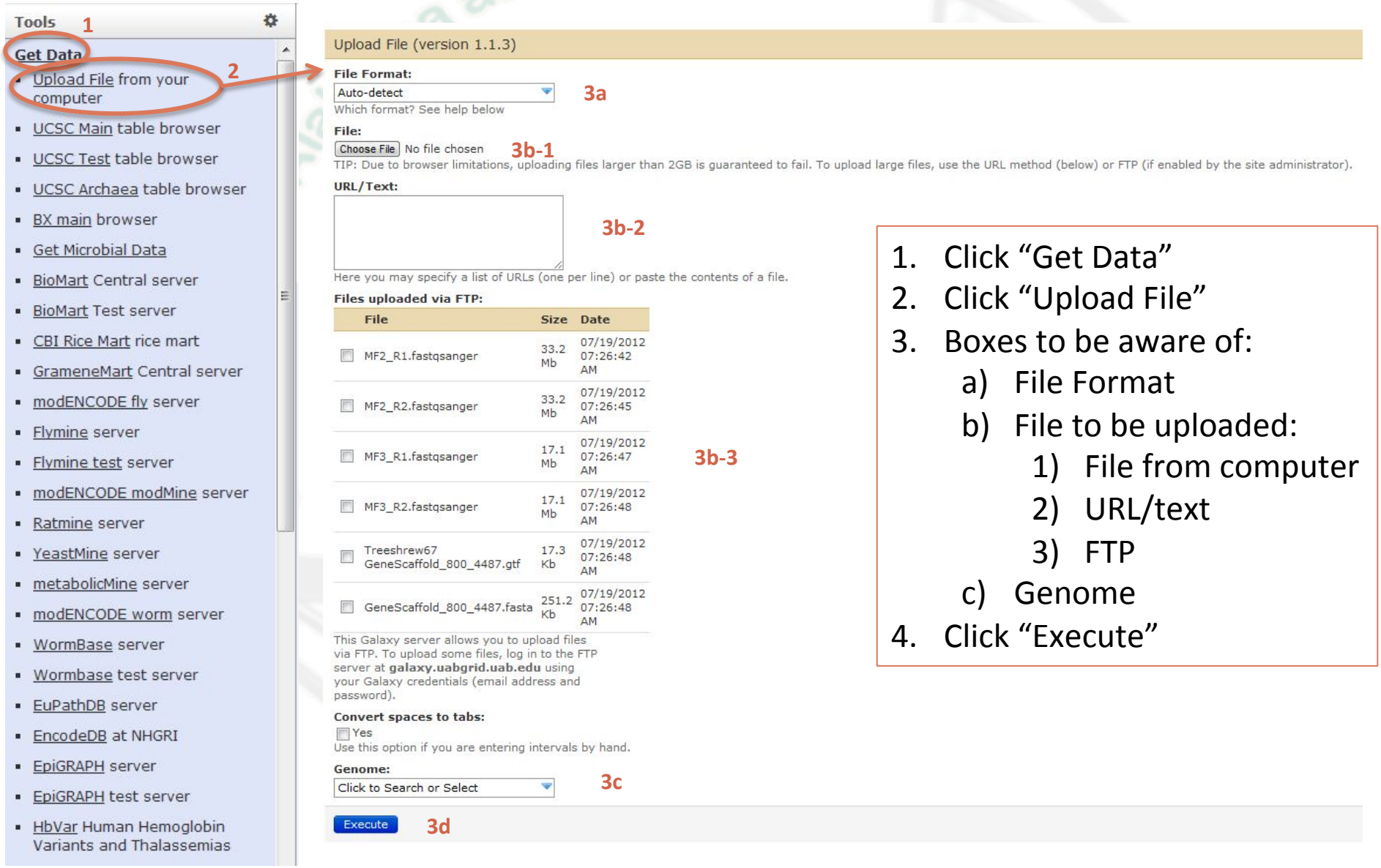
Cufflinks package

- **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

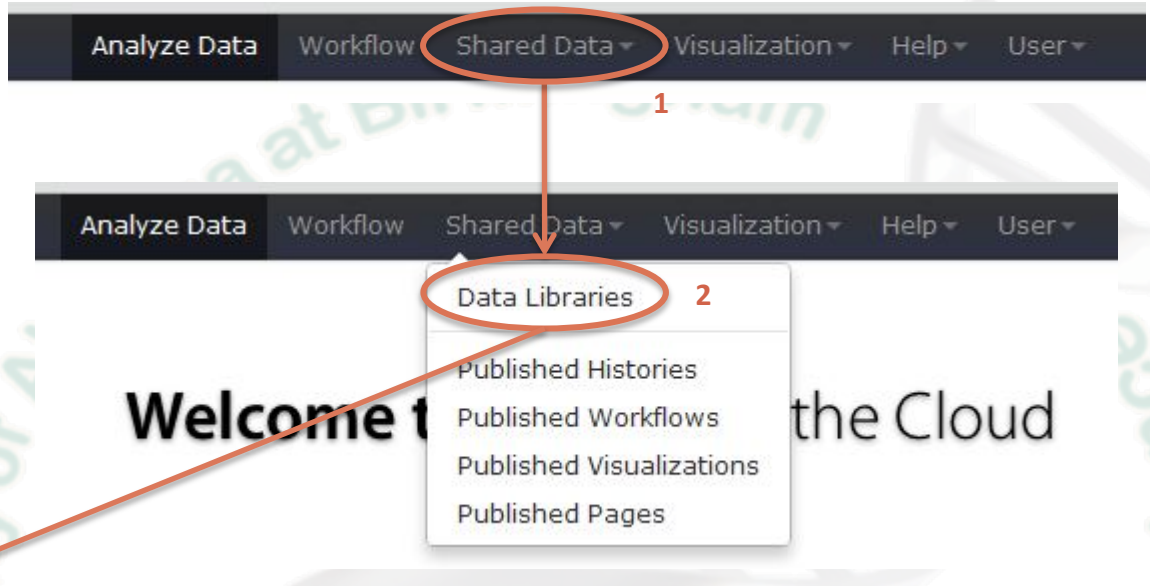
Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks Nature Protocols **7**, 562-578 (2012) doi:10.1038/nprot.2012.016

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



Data Library "GBS722-2014"

Raw data for various NGS demos that will be performed during class

Name	Message	Data type	Date uploaded	File size
RNA-Seq	Human data from a brain and adrenal sample			
Adrenal_1	None	fastqsanger	2014-02-14	7.8 MB
Adrenal_2	None	fastqsanger	2014-02-14	7.8 MB
Brain_1	None	fastqsanger	2014-02-14	5.9 MB
Brain_2	None	fastqsanger	2014-02-14	5.9 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 "GBS722-2014"

Raw data for various NGS demos that will be performed during class

Name	Message	Data type	Date uploaded	File size
RNA-Seq	Human data from a brain and adrenal sample			
Adrenal_1	None	fastqsanger	2014-02-14	7.8 MB
Adrenal_2	None	fastqsanger	2014-02-14	7.8 MB
Brain_1	None	fastqsanger	2014-02-14	5.9 MB
Brain_2	None	fastqsanger	2014-02-14	5.9 MB

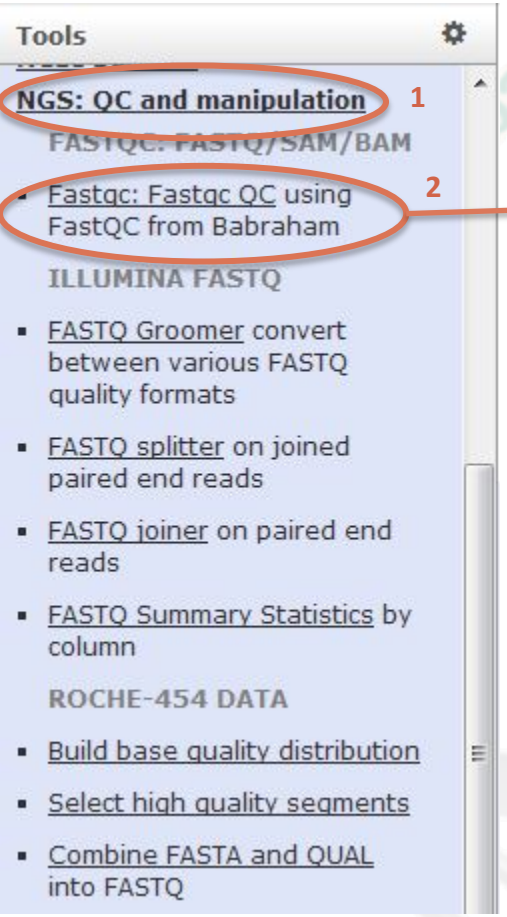
For selected datasets: Import to current history Go

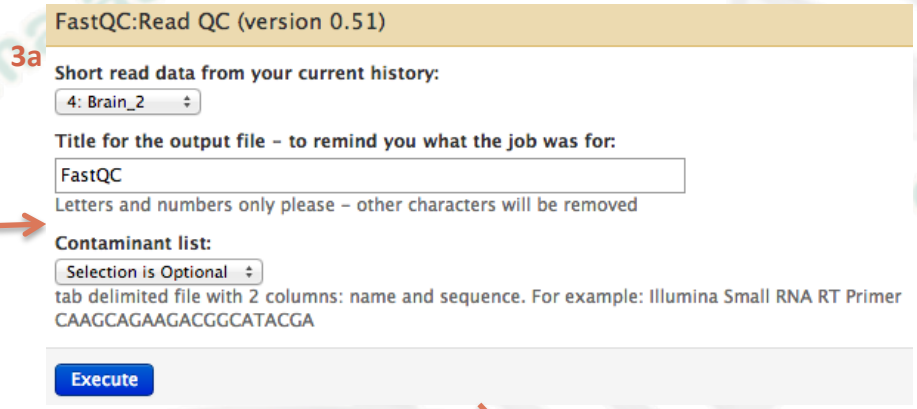
1

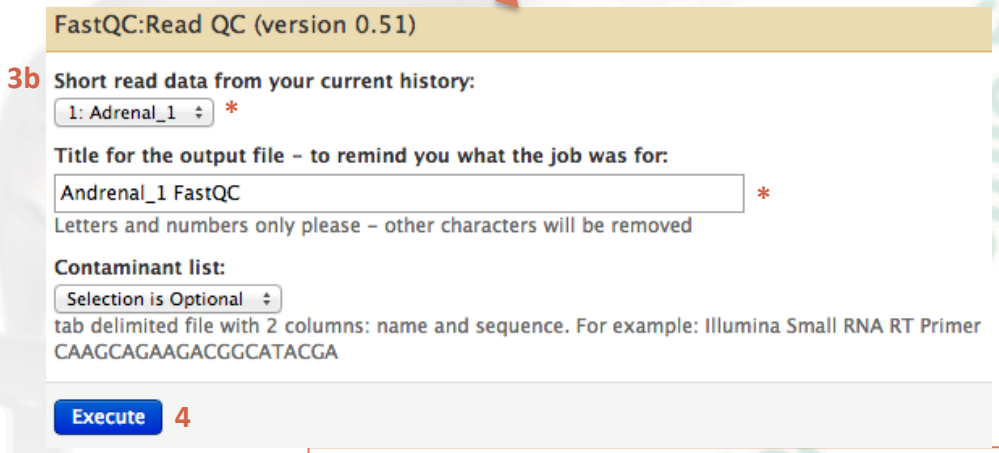
2

1. Check boxes of files you want to import
2. Choose "Import to current history" and then click "Go"

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

7: Adrenal2FastQC_Adrenal_2.html

6: Adrenal1FastQC_Adrenal_1.html

Adrenal_1 FastQC Report

FastQC Report
Sun 9 Feb 2014
Adrenal_1

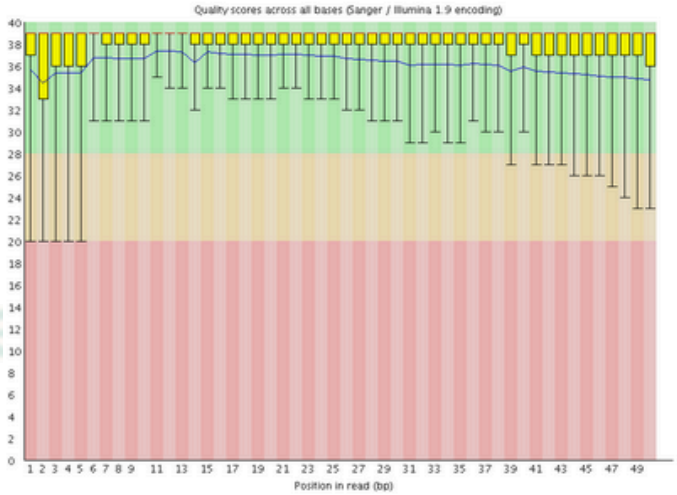
Summary

- 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

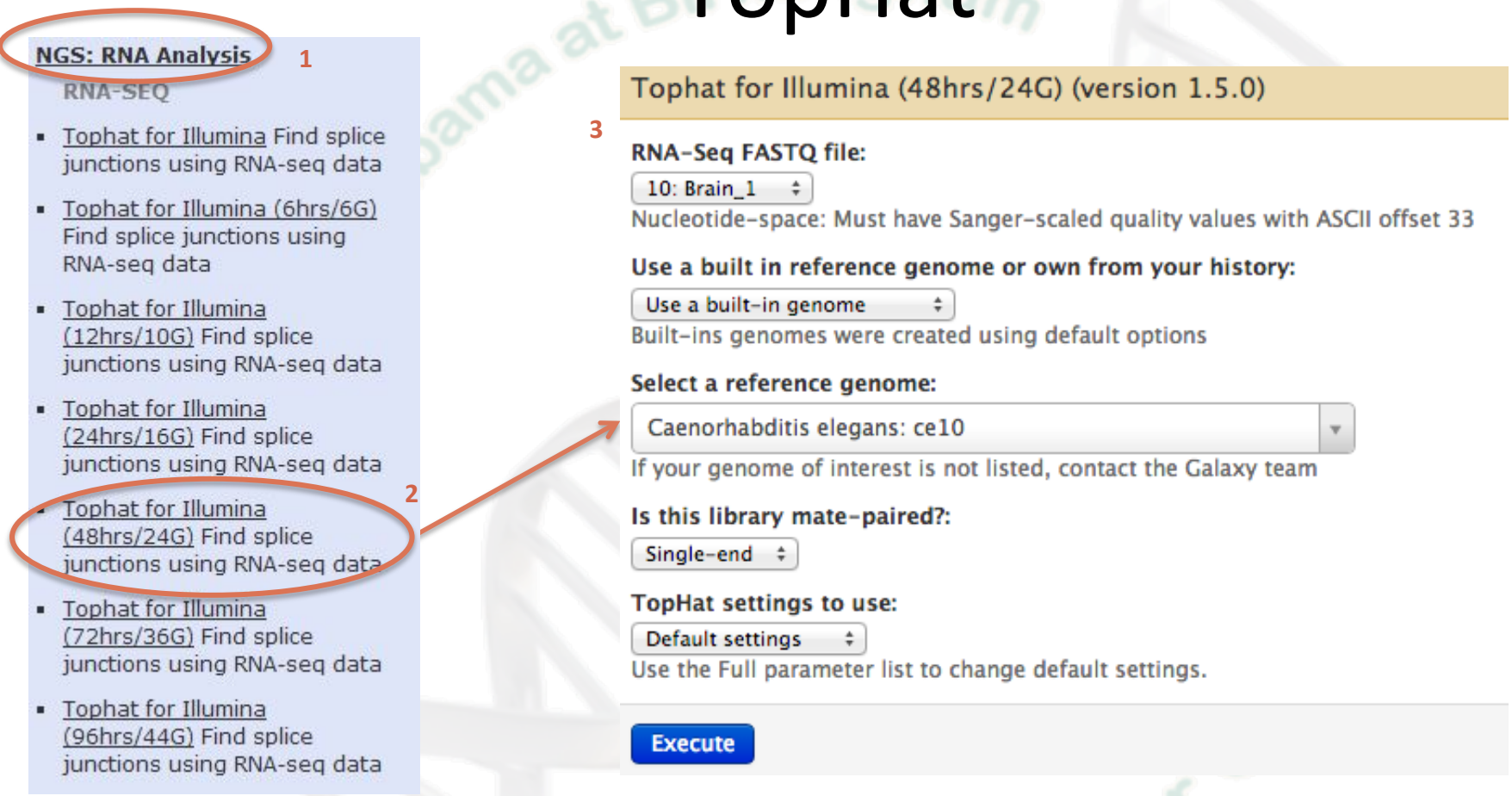
Basic Statistics

Measure	Value
Filename	Adrenal_1
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	50121
Filtered Sequences	0
Sequence length	50
%GC	58

Per base sequence quality

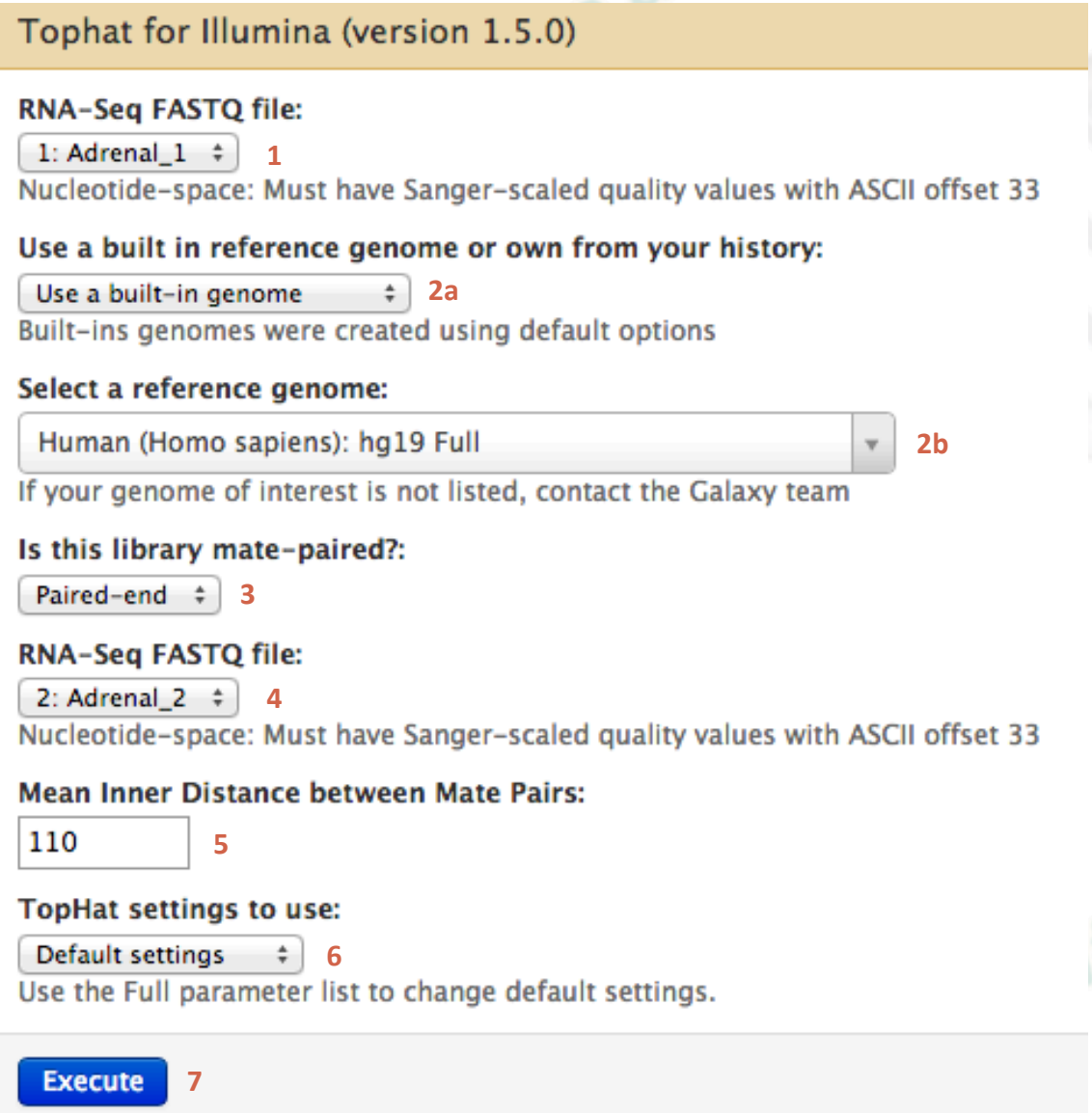


TopHat



1. Click on “NGS: RNA Analysis”
2. Click on “Tophat for Illumina (48hrs/24G)”
3. Default window with options appears

TopHat

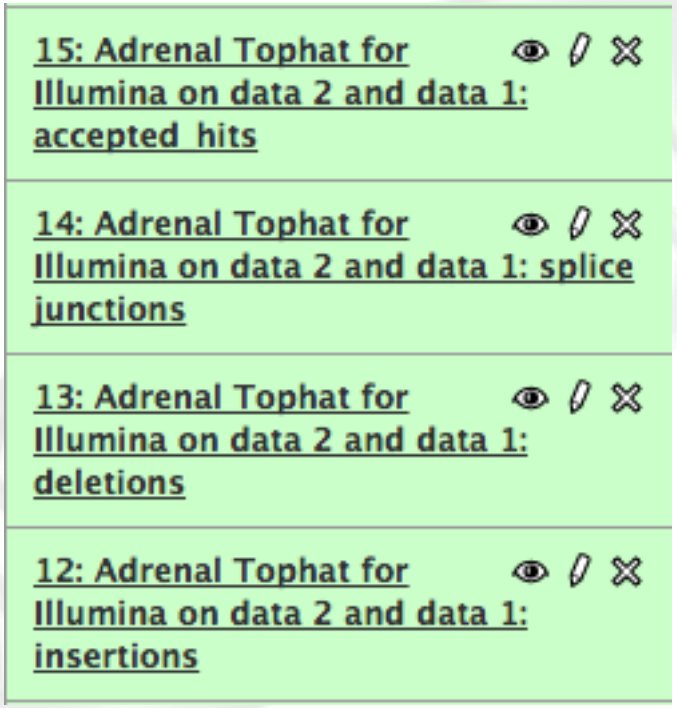


1. Select forward fastq read file
2. Select reference genome:
   a) Choose “Use a built-in genome”
   b) Select the reference genome
3. Select “Paired-end”
4. Select reverse fastq read file
5. Input “110” (ask sequencing center for this info)
6. Can choose “Commonly used” or “Full parameter list”
7. Click “Execute”
8. Do the exact same thing for the other sample

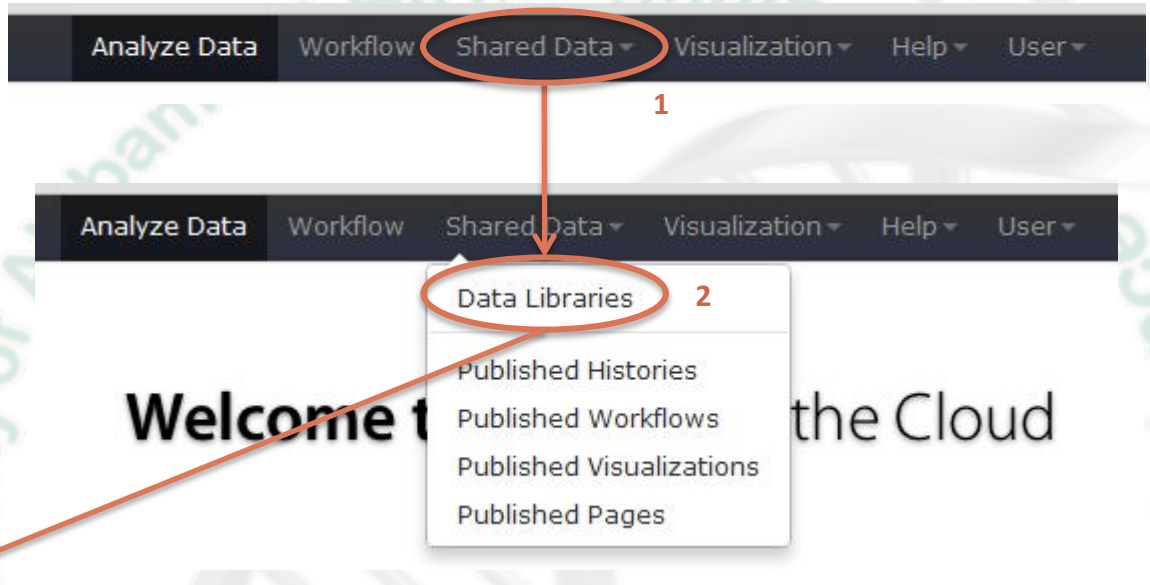
Note about FASTA files not already indexed in Galaxy

- 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



GTF Annotation Files



3

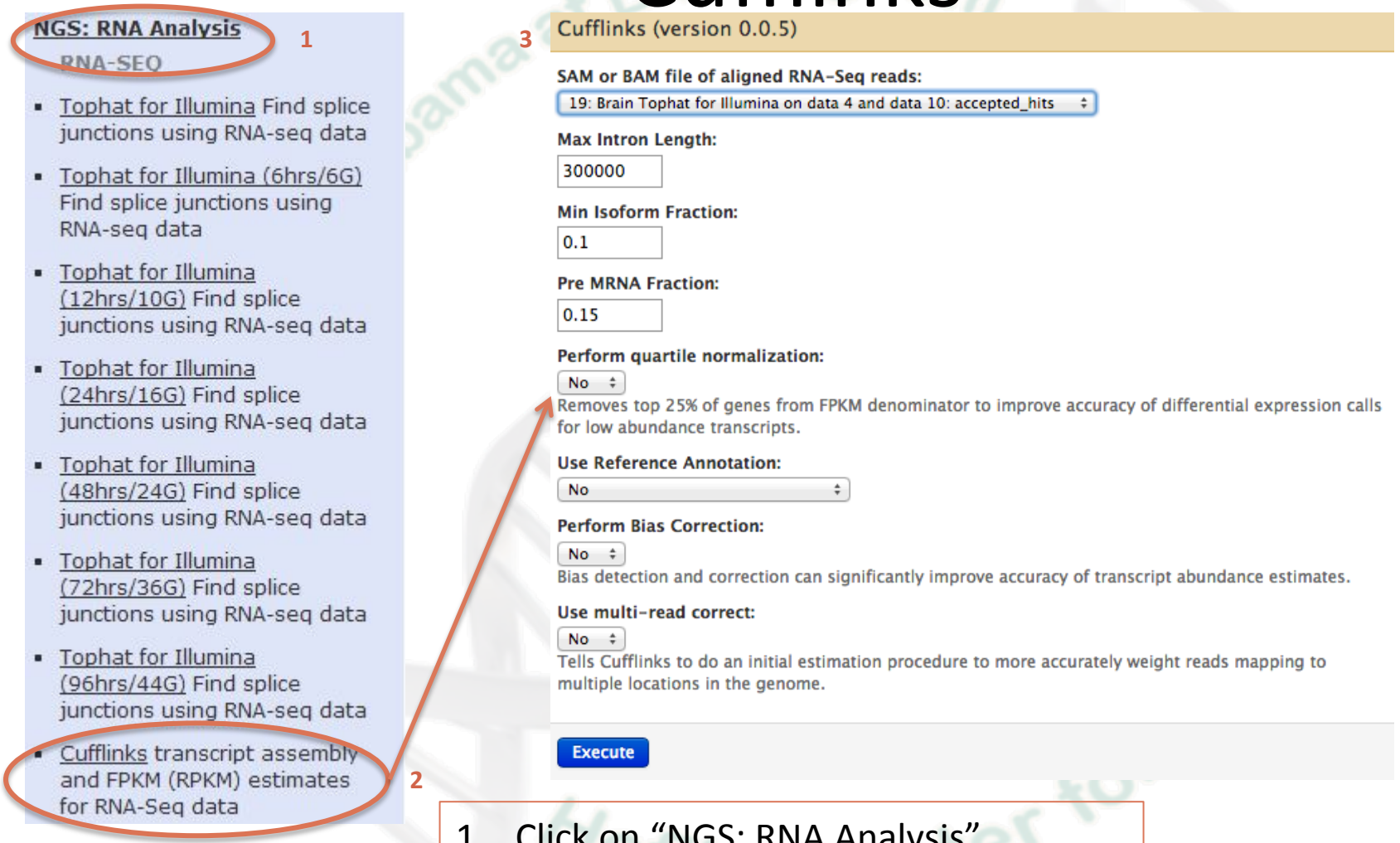
Data Library "Patched GTF annotation files for Cufflinks"

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

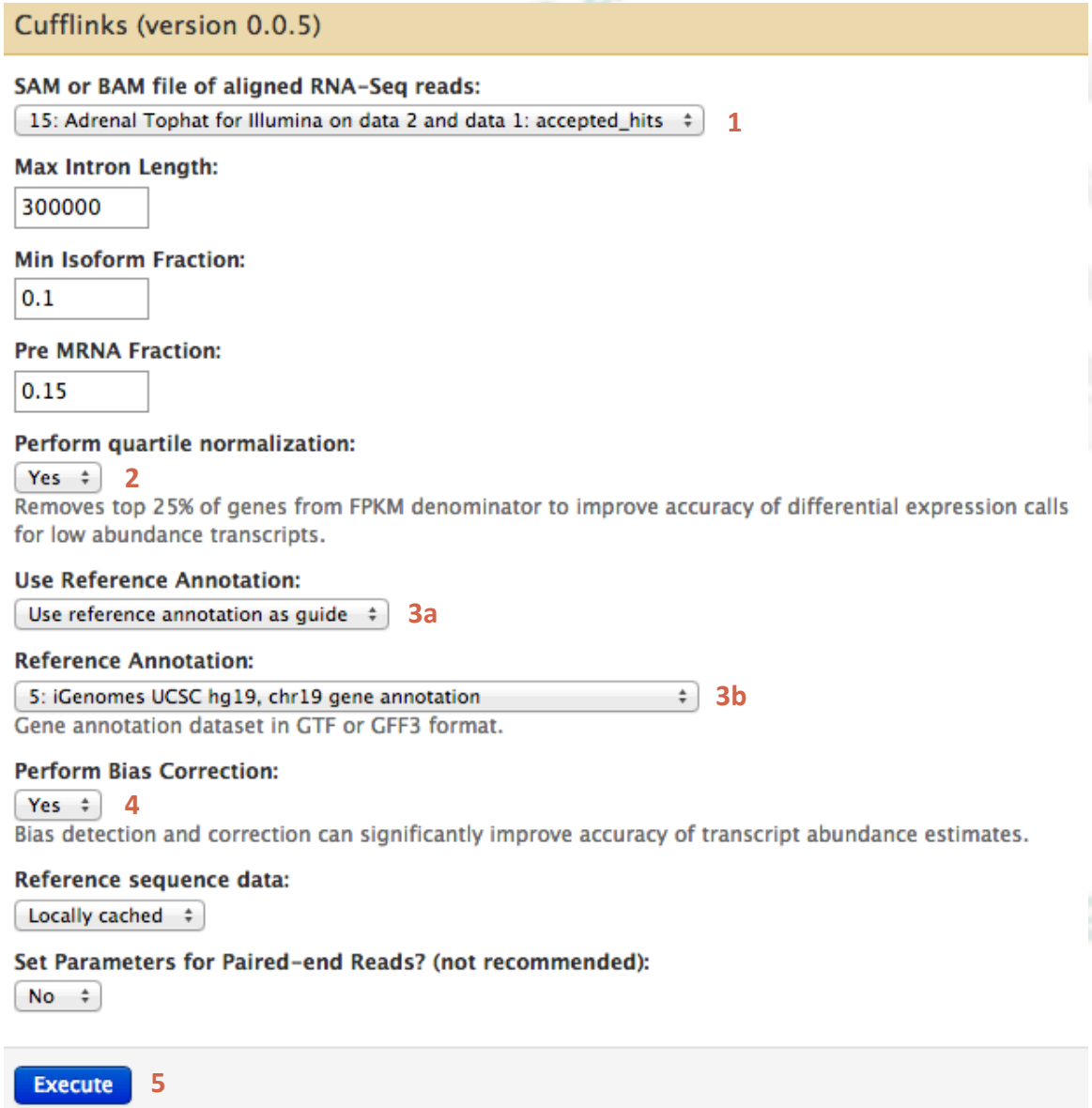
For selected datasets: Import to current history Go

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 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.

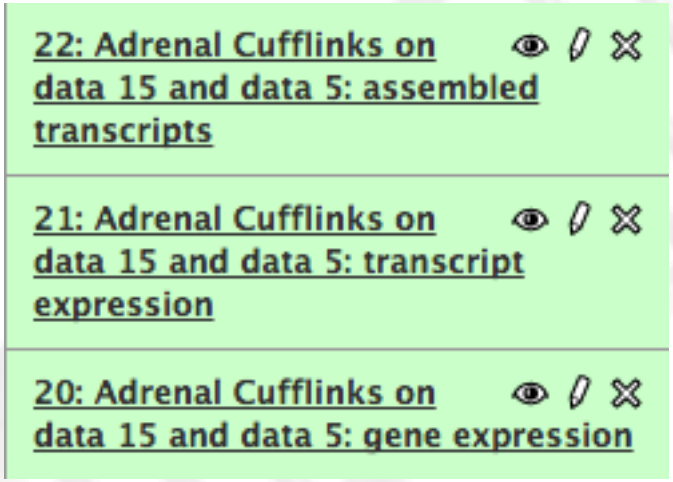
-F/–min-isoform-fraction <0.0-1.0>

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.

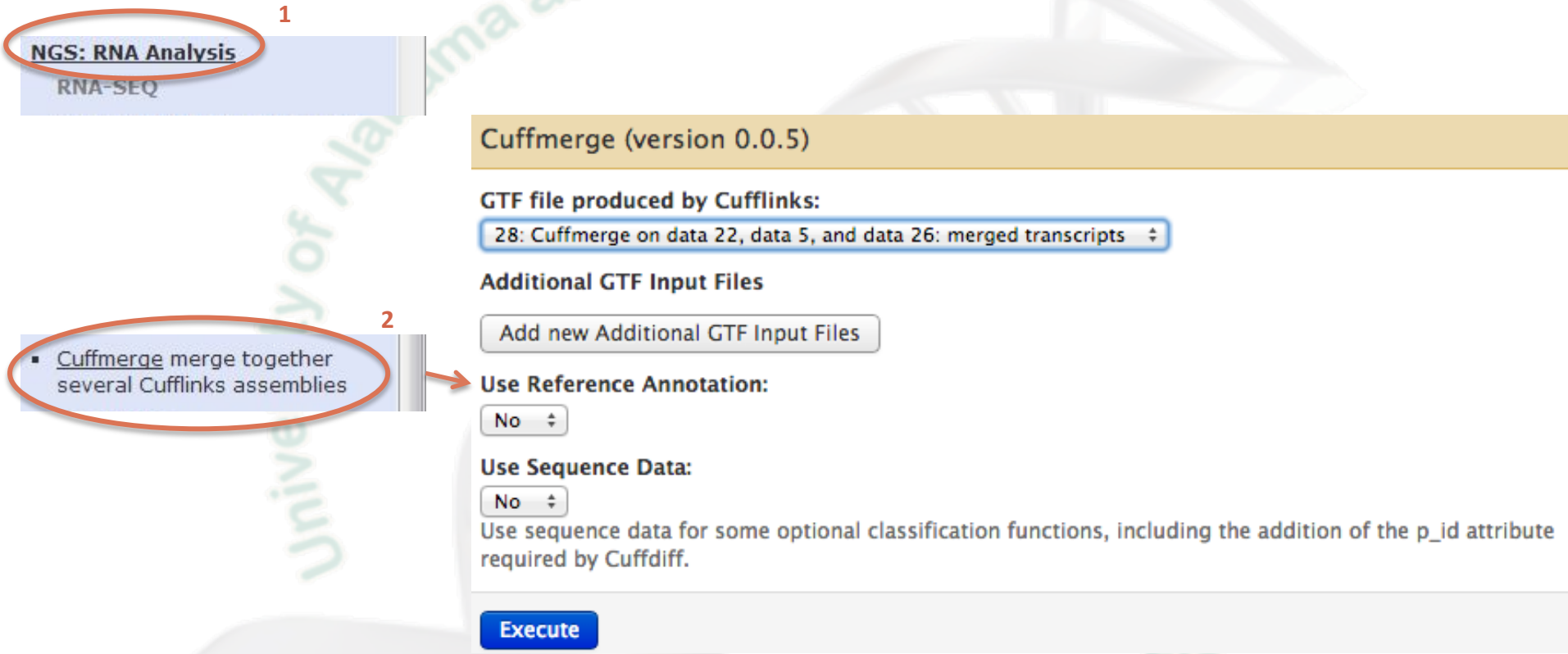
-j/–pre-mrna-fraction <0.0-1.0>

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

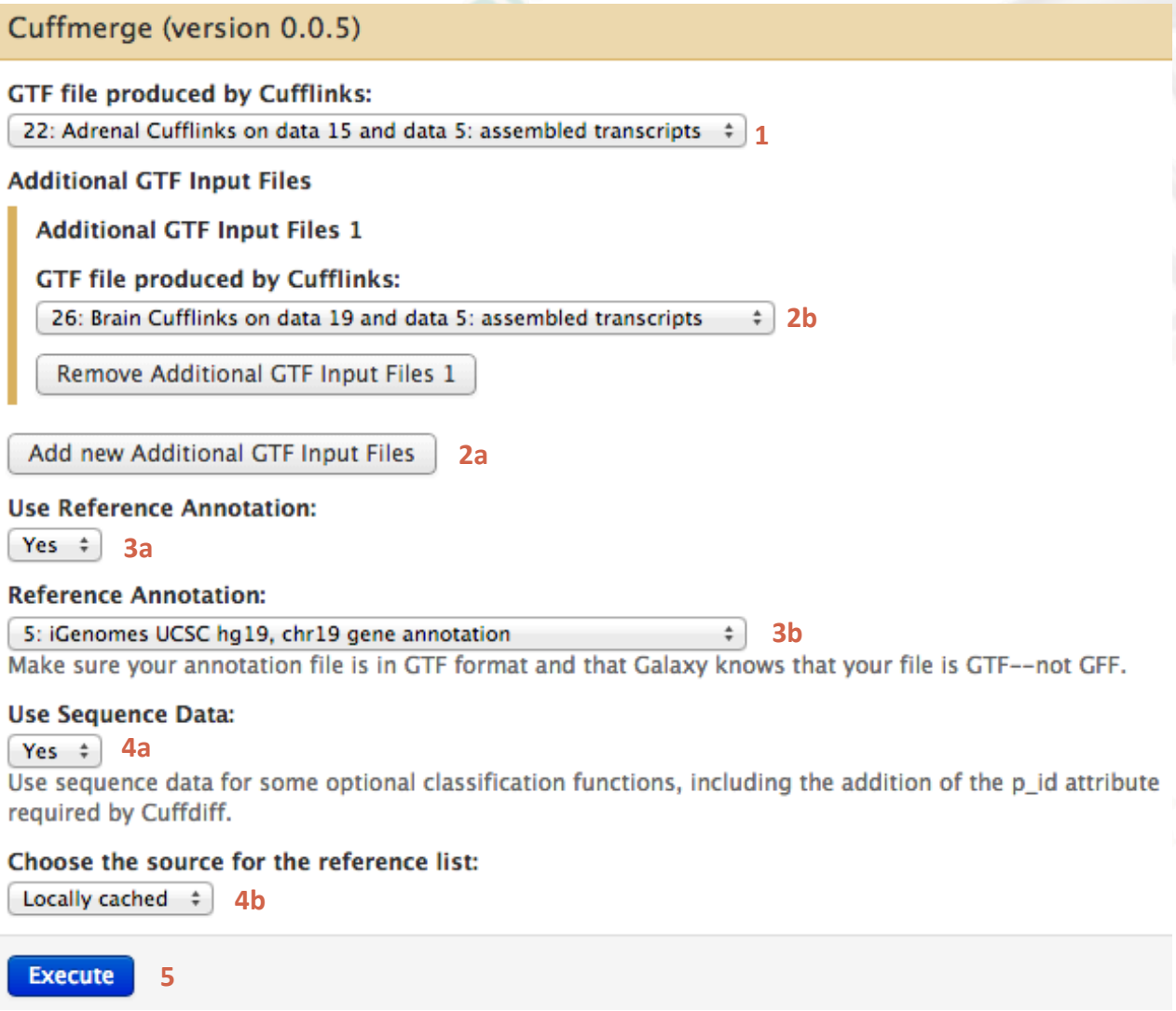


Cuffmerge



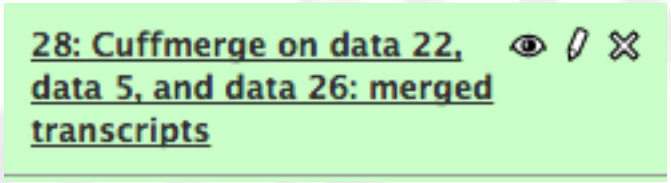
1. Click on “NGS: RNA Analysis”
2. Click on “Cuffmerge”
3. Default window with options appears

Cuffmerge

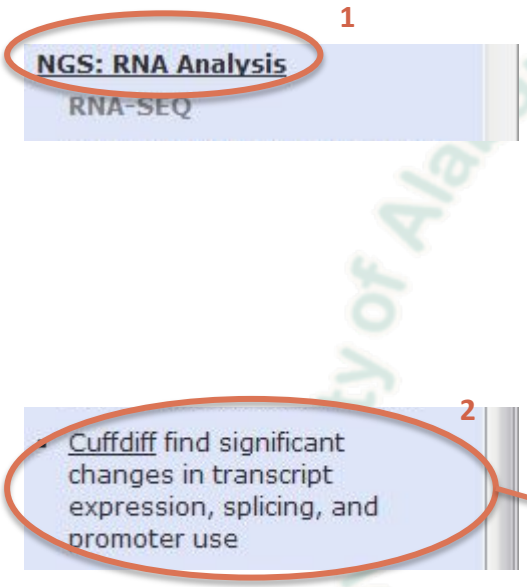


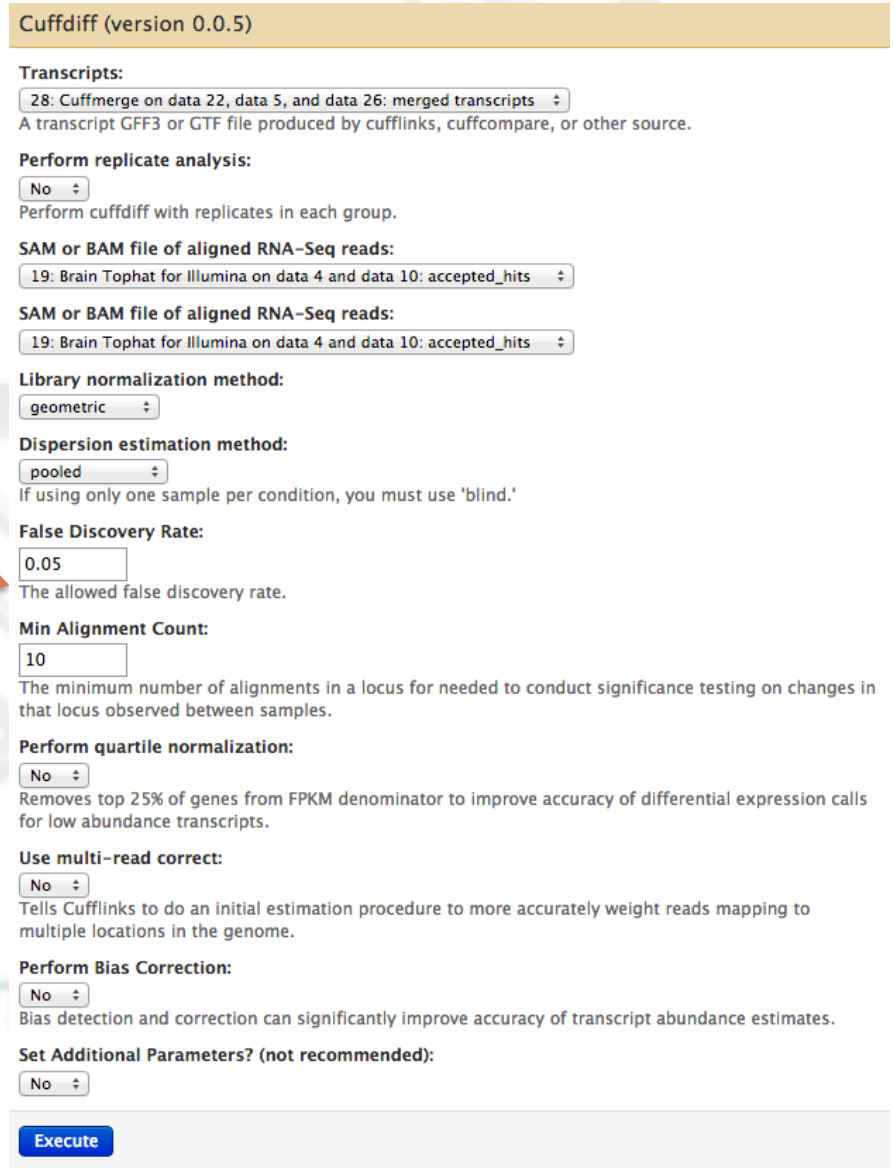
1. Choose GTF file produced by Cufflinks
2. Additional GTF Input Files:
    a) Click on “Add new Additional GTF Input Files”
    b) Choose other GTF file produced by Cufflinks
3. Reference Annotation:
    a) Select “Yes” to Use Reference Annotation
    b) Choose GTF Reference Annotation file from history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

Cuffmerge output files



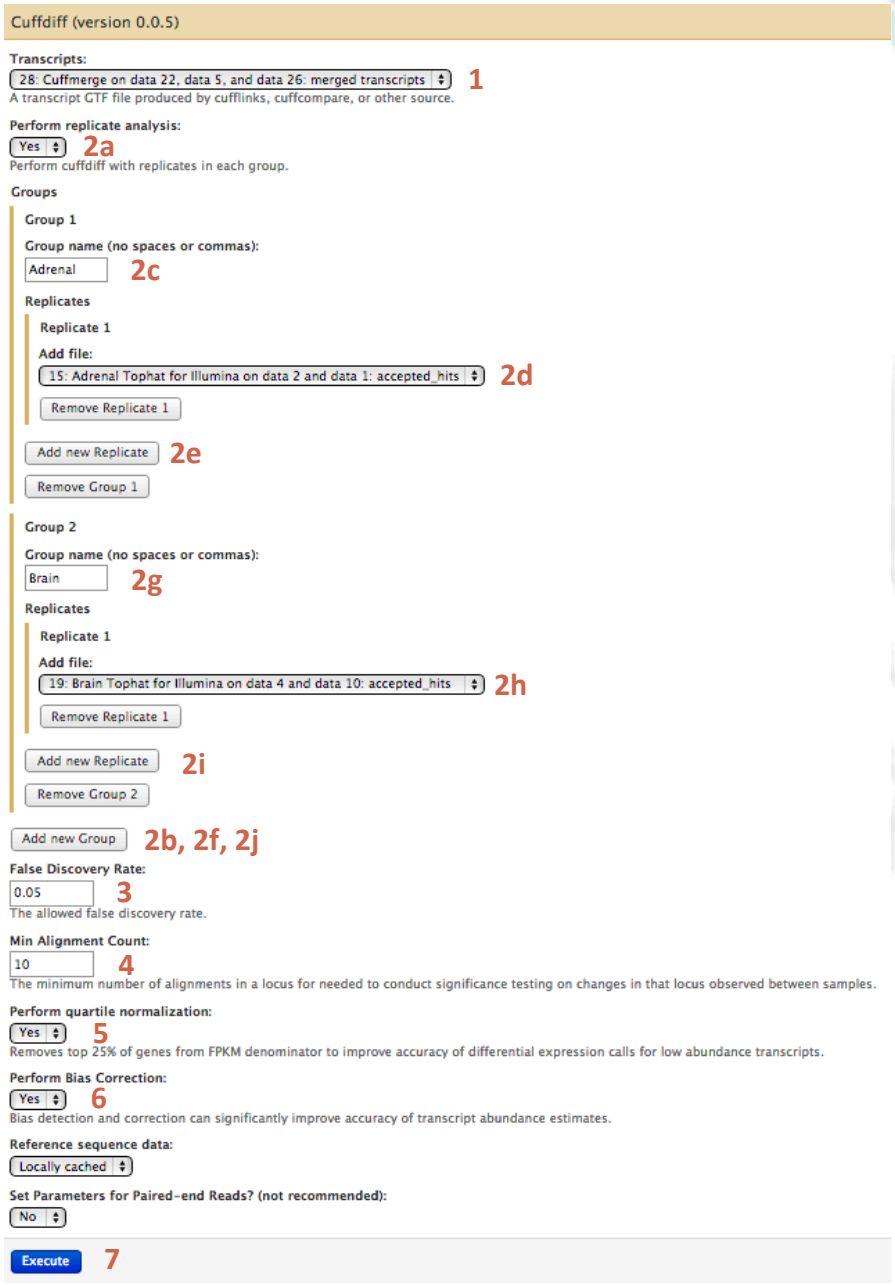
Cuffdiff





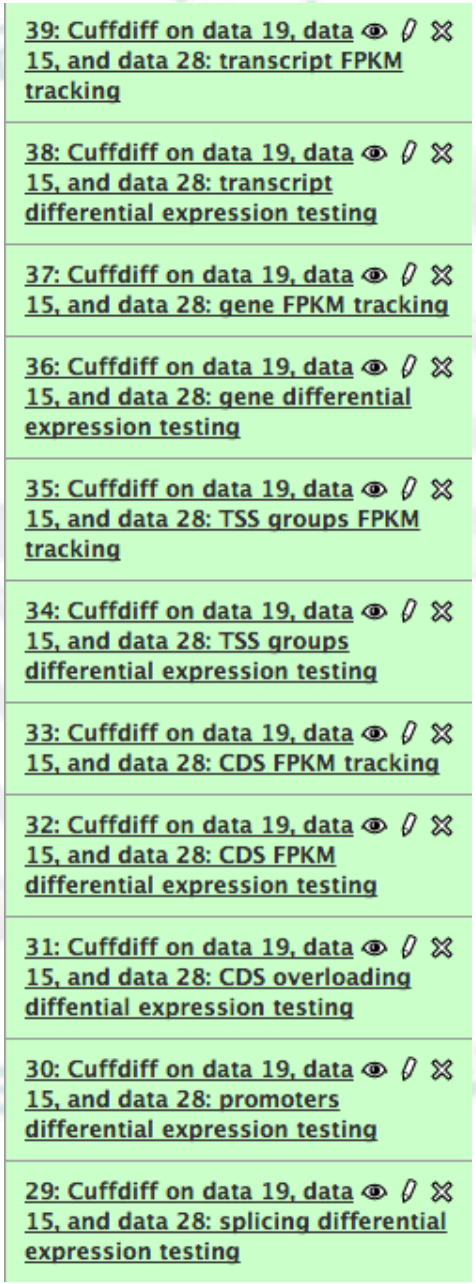
1. Click on “NGS: RNA Analysis”
2. Click on “Cuffdiff”
3. Default window with options appears

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



Transcript differential expression testing output

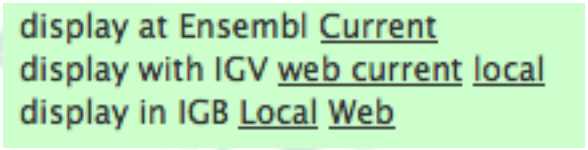
test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
TCONS_00000001	XLOC_000001	OR4F5	chr1:69090-70008	Control	Treated	NOTEST	0	7.91888	1.79769e+308	1.79769e+308	0.369441	1	no
TCONS_00000002	XLOC_000002	LOC100132062	chr1:323891-328581	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

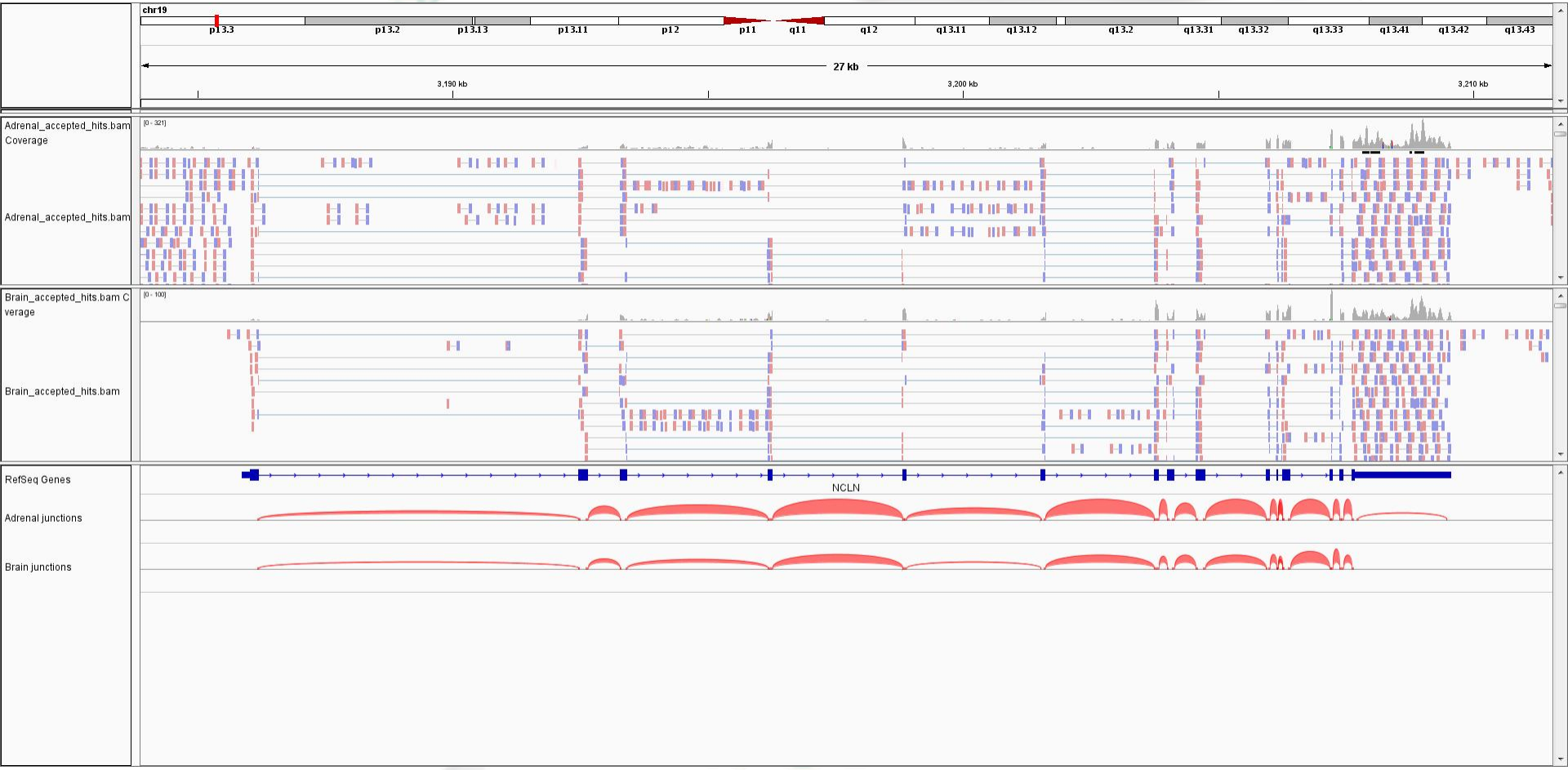
- http://www.broadinstitute.org/igv/
- Several ways to view the accepted_hits.bam file from TopHat:
  - Download the bam file to your computer (don't forget to download the bam_index file (*.bai) and then load into IGV
  - View them directly from Galaxy (no downloading required)



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
[0 - 60]
Zoom in to see alignments.
Brain_accepted_hits.bam Coverage
Brain_accepted_hits.bam
[0 - 60]
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

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/