Approaches to Bioinformatic Data Analysis
RNA-Seq Analysis using Galaxy

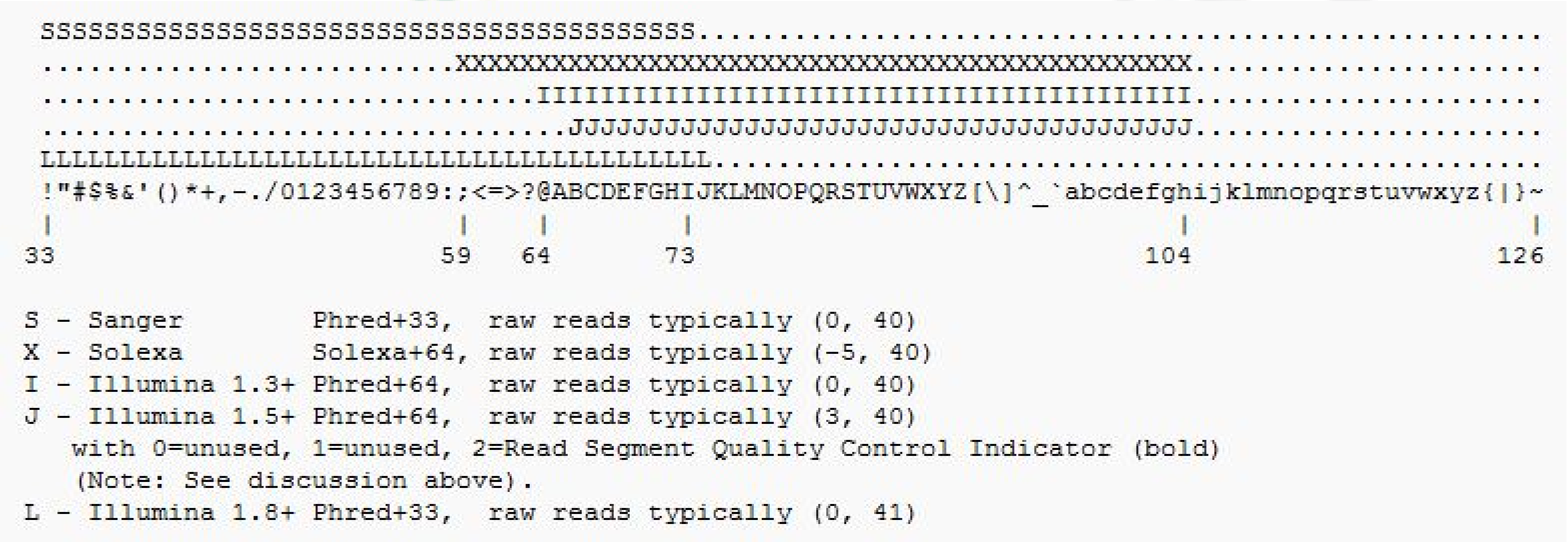
David Crossman, Ph.D.
UAB Heflin Center for Genomic Science

Immersion Course
Friday, April 5, 2013

Contents

- FASTQ anatomy
- RNA-Seq demo
- Genomics Viewer (IGV) demo
- Whole Genome/Exome demo
- References and web links

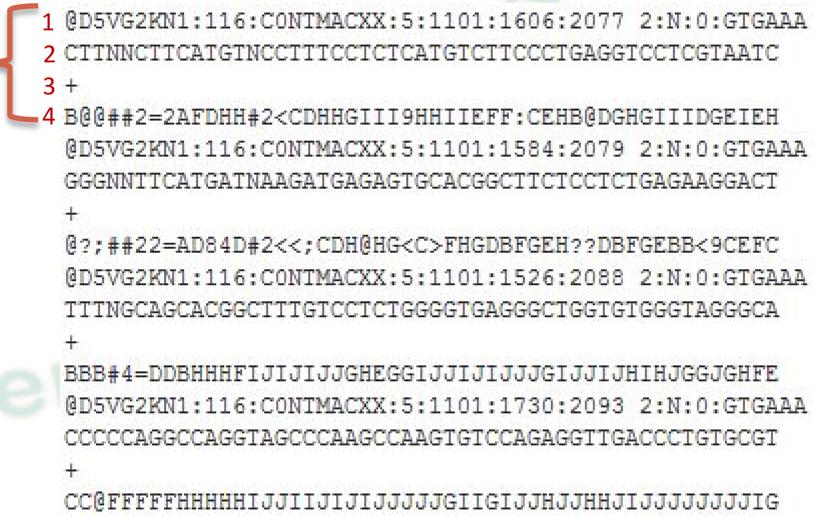
NGS FASTQ file format



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL....................................................
 !"#$%&'()*+,-./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…

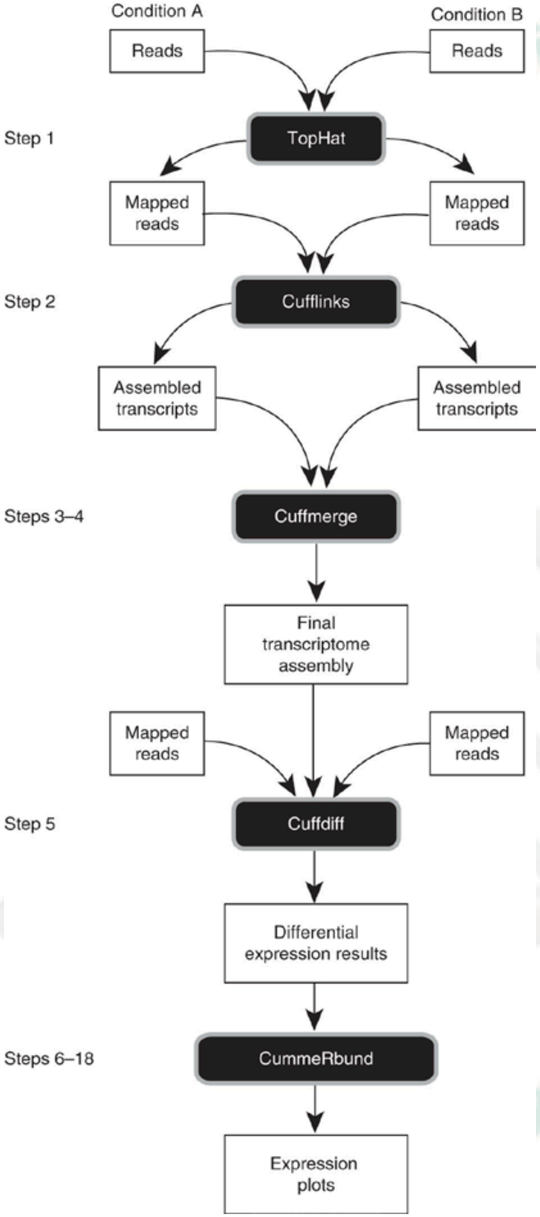


```
1 @D5VG2KN1:116:C0NTMACXX:5:1101:1606:2077 2:N:0:GTGAAA
2 CTTNNCTTCATGTNCCTTTCCTCTCATGTCTTCCCTGAGGTCCTCGTAATC
3 +
4 B@@##2=2AFDHH#2<CDHHGIII9HHIIEFF:CEHB@DGHGIIIDGEIEH
  @D5VG2KN1:116:C0NTMACXX:5:1101:1584:2079 2:N:0:GTGAAA
  GGGNNTTCATGATNAAGATGAGAGTGCACGGCTTCTCCTCTGAGAAGGACT
  +
  @?;##22=AD84D#2<<;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
```

Contents

- FASTQ anatomy
- RNA-Seq demo
- Genomics Viewer (IGV) demo
- Whole Genome/Exome demo
- References and web links

RNA-Seq pipeline



Condition A

Reads

Condition B

Reads

Step 1

TopHat

Mapped reads

Mapped reads

Step 2

Cufflinks

Assembled transcripts

Assembled transcripts

Steps 3–4

Cuffmerge

Final transcriptome assembly

Mapped reads

Mapped reads

Step 5

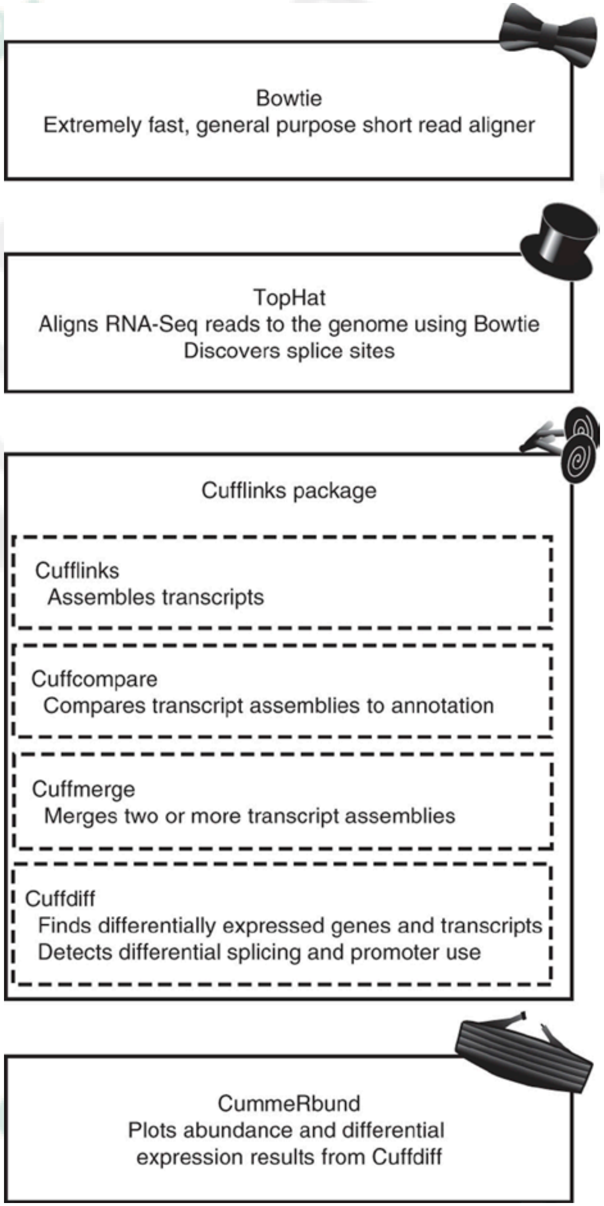
Cuffdiff

Differential expression results

Steps 6–18

CummeRbund

Expression plots



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

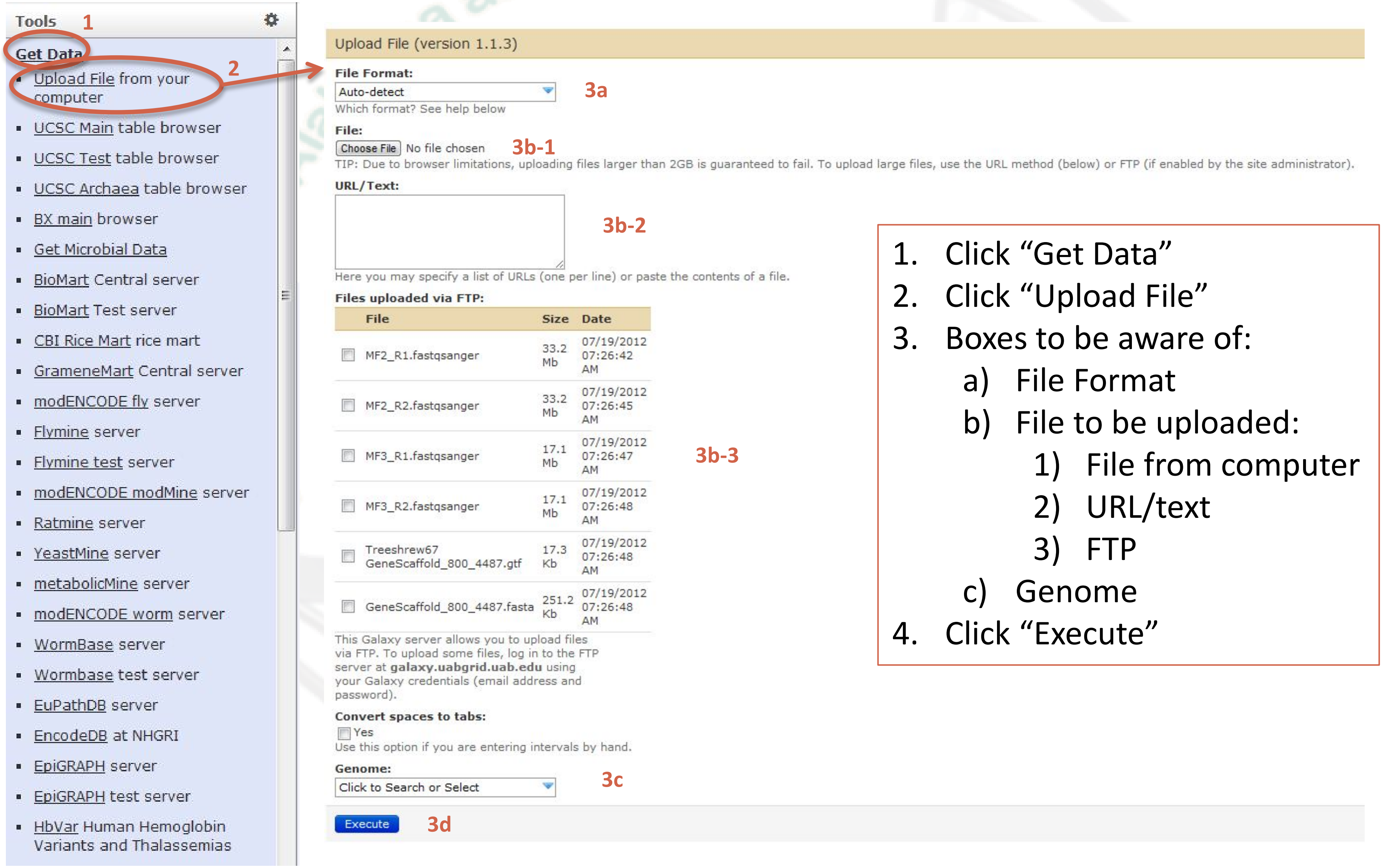
Cuffmerge
Merges two or more transcript assemblies

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



Tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Test table browser
- UCSC Archaea table browser
- BX main browser
- Get Microbial Data
- BioMart Central server
- BioMart Test server
- CBI Rice Mart rice mart
- GrameneMart Central server
- modENCODE fly server
- Flymine server
- Flymine test server
- modENCODE modMine server
- Ratmine server
- YeastMine server
- metabolicMine server
- modENCODE worm server
- WormBase server
- Wormbase test server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server
- EpiGRAPH test server
- HbVar Human Hemoglobin Variants and Thalassemias

Upload File (version 1.1.3)

File Format:
Auto-detect
Which format? See help below

File:
Choose File No file chosen
TIP: Due to browser limitations, uploading files larger than 2GB is guaranteed to fail. To upload large files, use the URL method (below) or FTP (if enabled by the site administrator).

URL/Text:

Here you may specify a list of URLs (one per line) or paste the contents of a file.

Files uploaded via FTP:

	File	Size	Date
☐	MF2_R1.fastqsanger	33.2 Mb	07/19/2012 07:26:42 AM
☐	MF2_R2.fastqsanger	33.2 Mb	07/19/2012 07:26:45 AM
☐	MF3_R1.fastqsanger	17.1 Mb	07/19/2012 07:26:47 AM
☐	MF3_R2.fastqsanger	17.1 Mb	07/19/2012 07:26:48 AM
☐	Treeshrew67 GeneScaffold_800_4487.gtf	17.3 Kb	07/19/2012 07:26:48 AM
☐	GeneScaffold_800_4487.fasta	251.2 Kb	07/19/2012 07:26:48 AM

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at **galaxy.uabgrid.uab.edu** using your Galaxy credentials (email address and password).

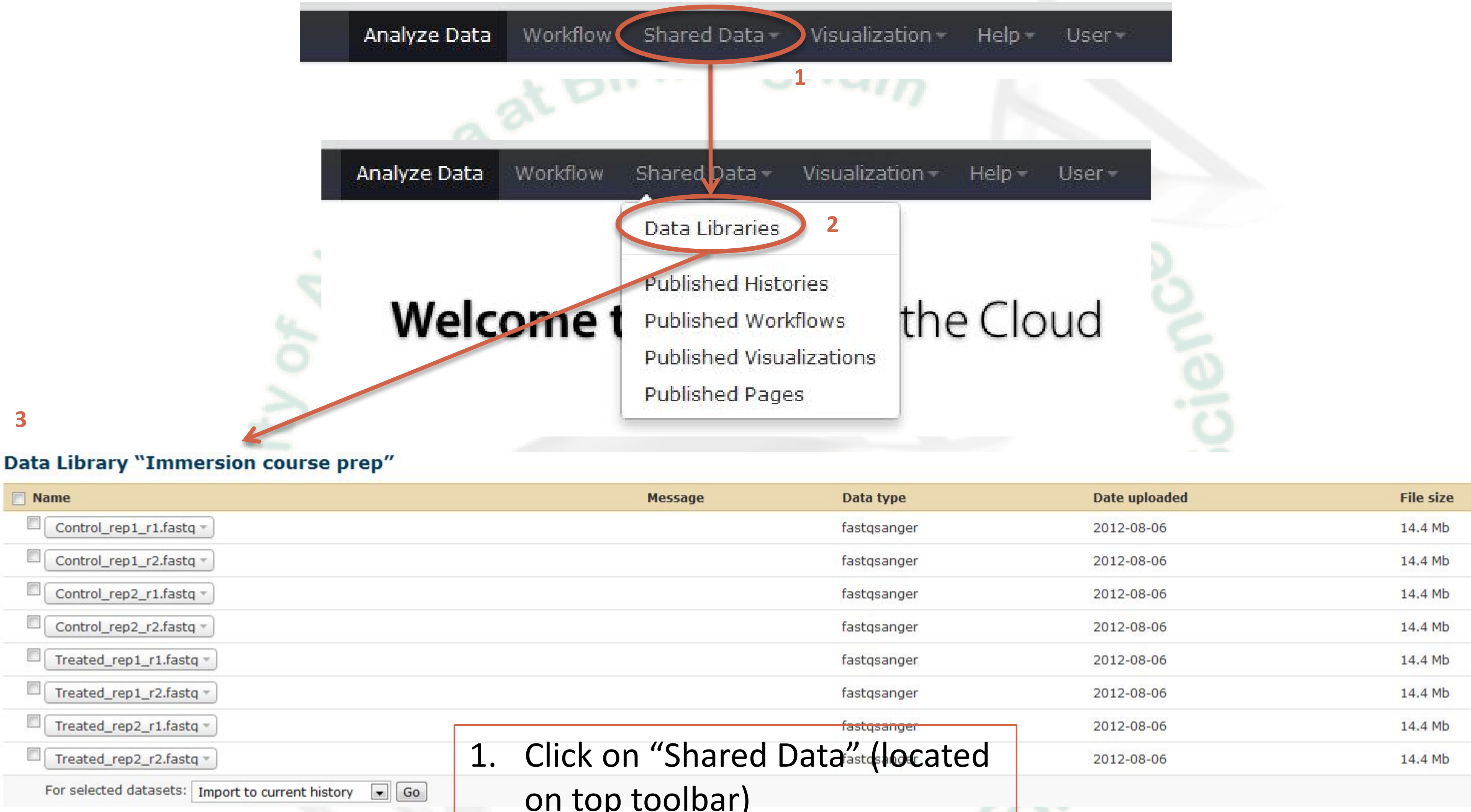
Convert spaces to tabs:
☐ Yes
Use this option if you are entering intervals by hand.

Genome:
Click to Search or Select

Execute

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



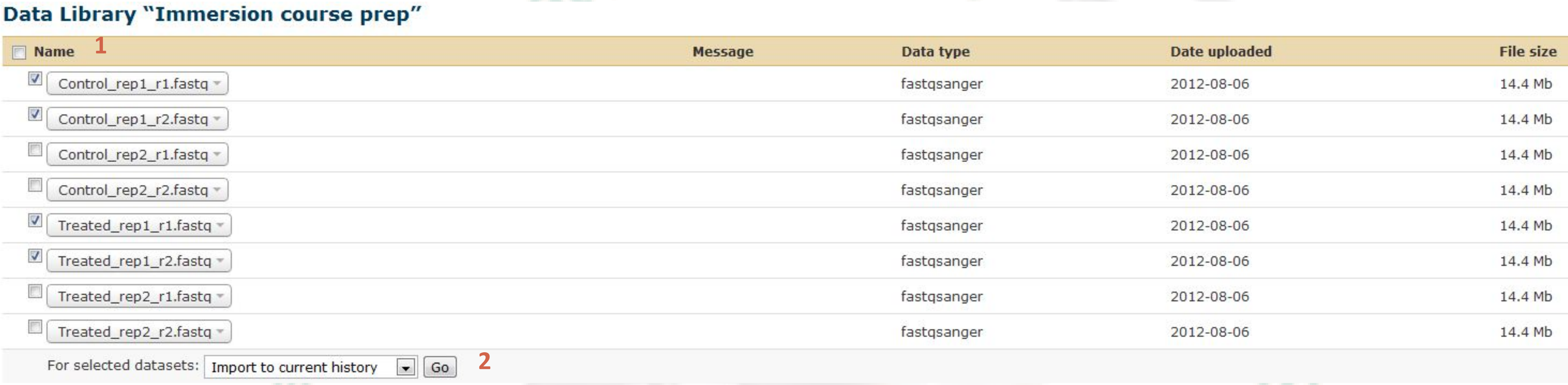
Data Library "Immersion course prep"

□ 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"

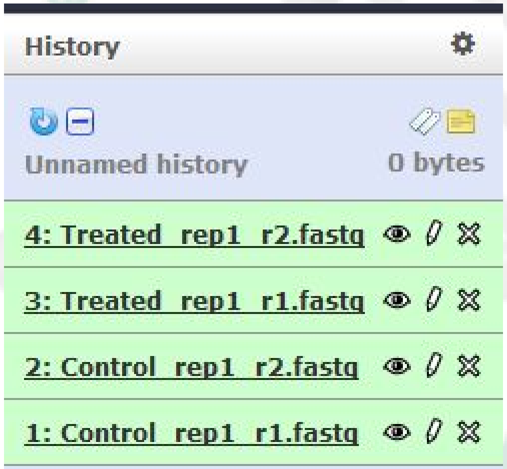
1

□ 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

2

3



History

Unnamed history 0 bytes

4: Treated_rep1_r2.fastq

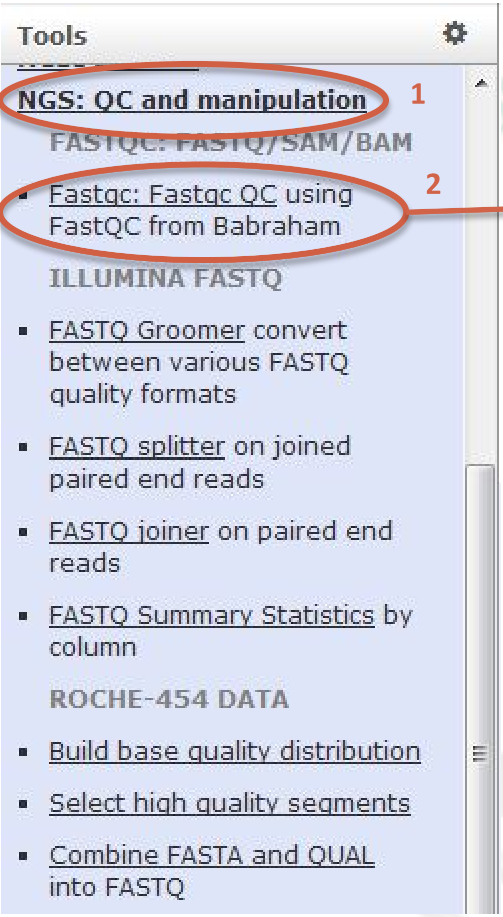
3: Treated_rep1_r1.fastq

2: Control_rep1_r2.fastq

1: Control_rep1_r1.fastq

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



Tools

NGS: QC and manipulation

FASTQC: FASTQ/SAM/BAM

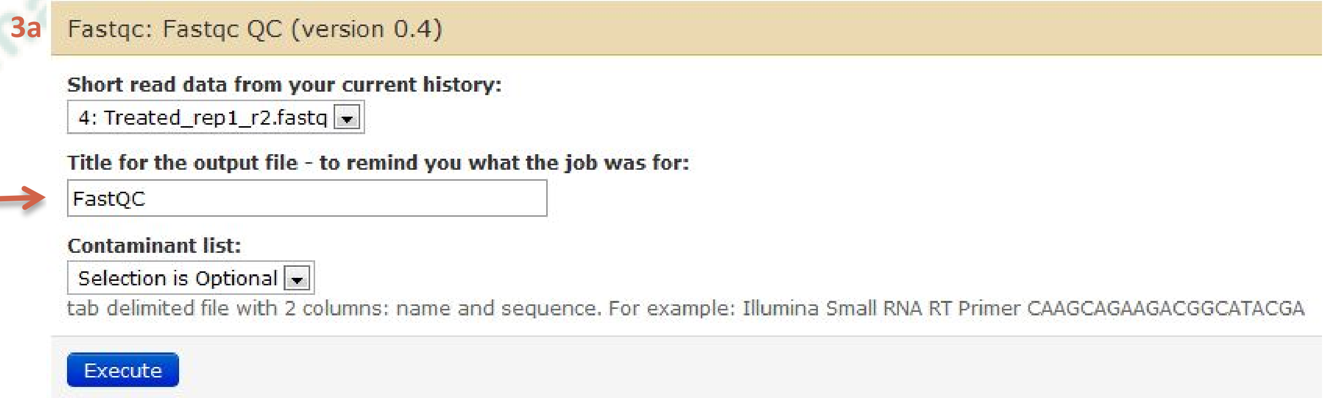
- Fastqc: Fastqc QC using FastQC from Babraham

ILLUMINA FASTQ

- FASTQ Groomer convert between various FASTQ quality formats
- FASTQ splitter on joined paired end reads
- FASTQ joiner on paired end reads
- FASTQ Summary Statistics by column

ROCHE-454 DATA

- Build base quality distribution
- Select high quality segments
- Combine FASTA and QUAL into FASTQ

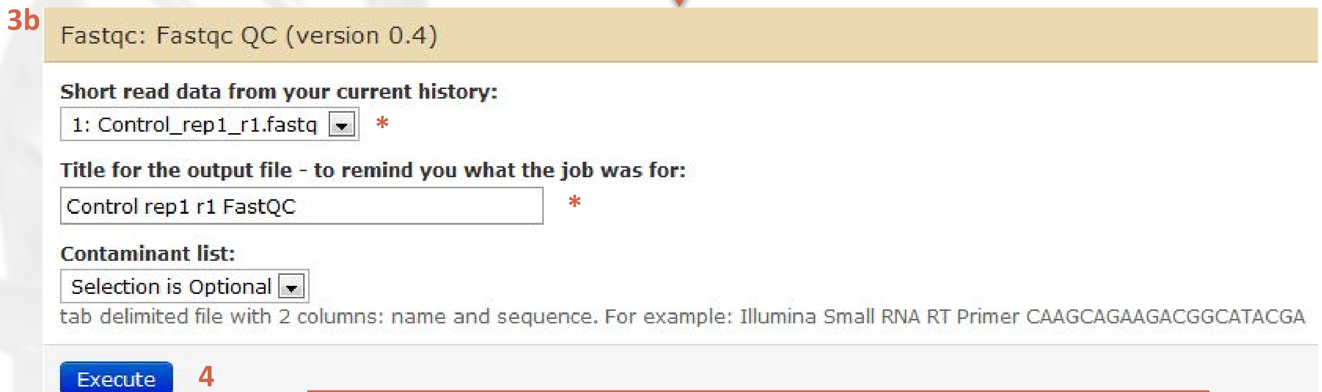


Fastqc: Fastqc QC (version 0.4)

Short read data from your current history:
4: Treated_rep1_r2.fastq

Title for the output file - to remind you what the job was for:
FastQC

Contaminant list:
Selection is Optional
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA



Fastqc: Fastqc QC (version 0.4)

Short read data from your current history:
1: Control_rep1_r1.fastq *

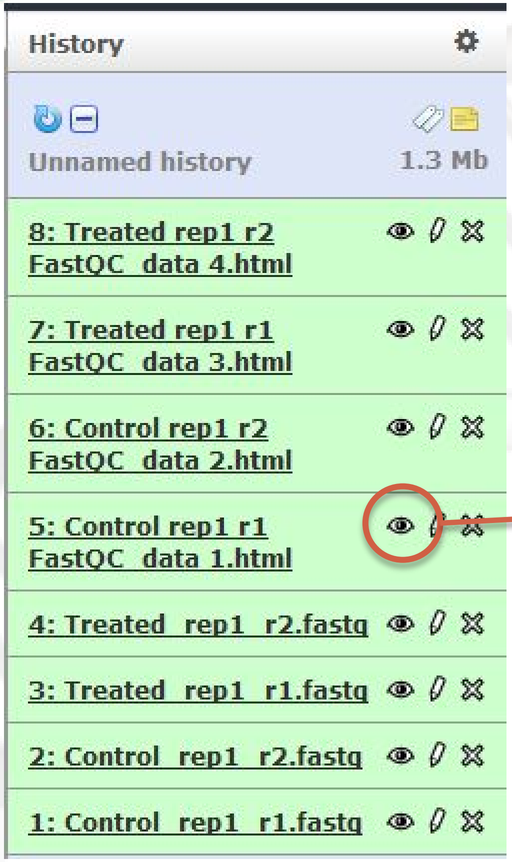
Title for the output file - to remind you what the job was for:
Control rep1 r1 FastQC *

Contaminant list:
Selection is Optional
tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAGACGGCATACGA

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!



History

Unnamed history 1.3 Mb

8: Treated rep1 r2 FastQC_data 4.html

7: Treated rep1 r1 FastQC_data 3.html

6: Control rep1 r2 FastQC_data 2.html

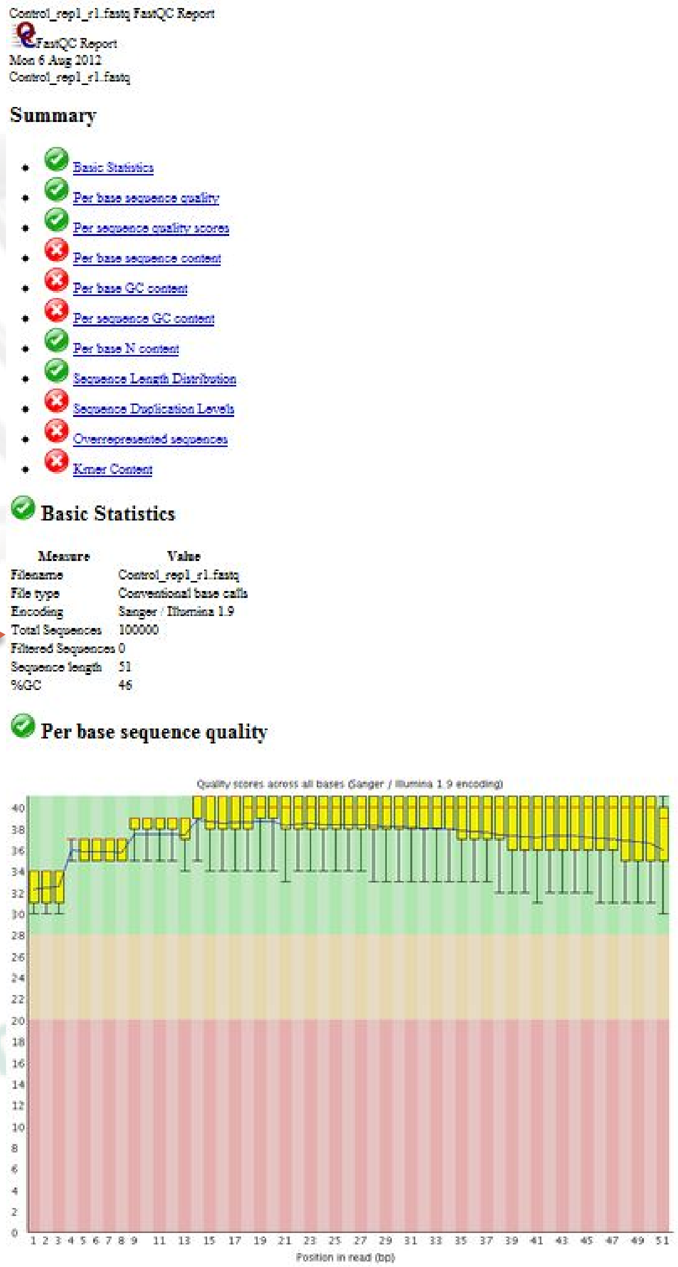
5: Control rep1 r1 FastQC_data 1.html

4: Treated_rep1_r2.fastq

3: Treated_rep1_r1.fastq

2: Control_rep1_r2.fastq

1: Control_rep1_r1.fastq



Control_rep1_r1.fastq FastQC Report

FastQC Report
Mon 6 Aug 2012
Control_rep1_r1.fastq

Summary

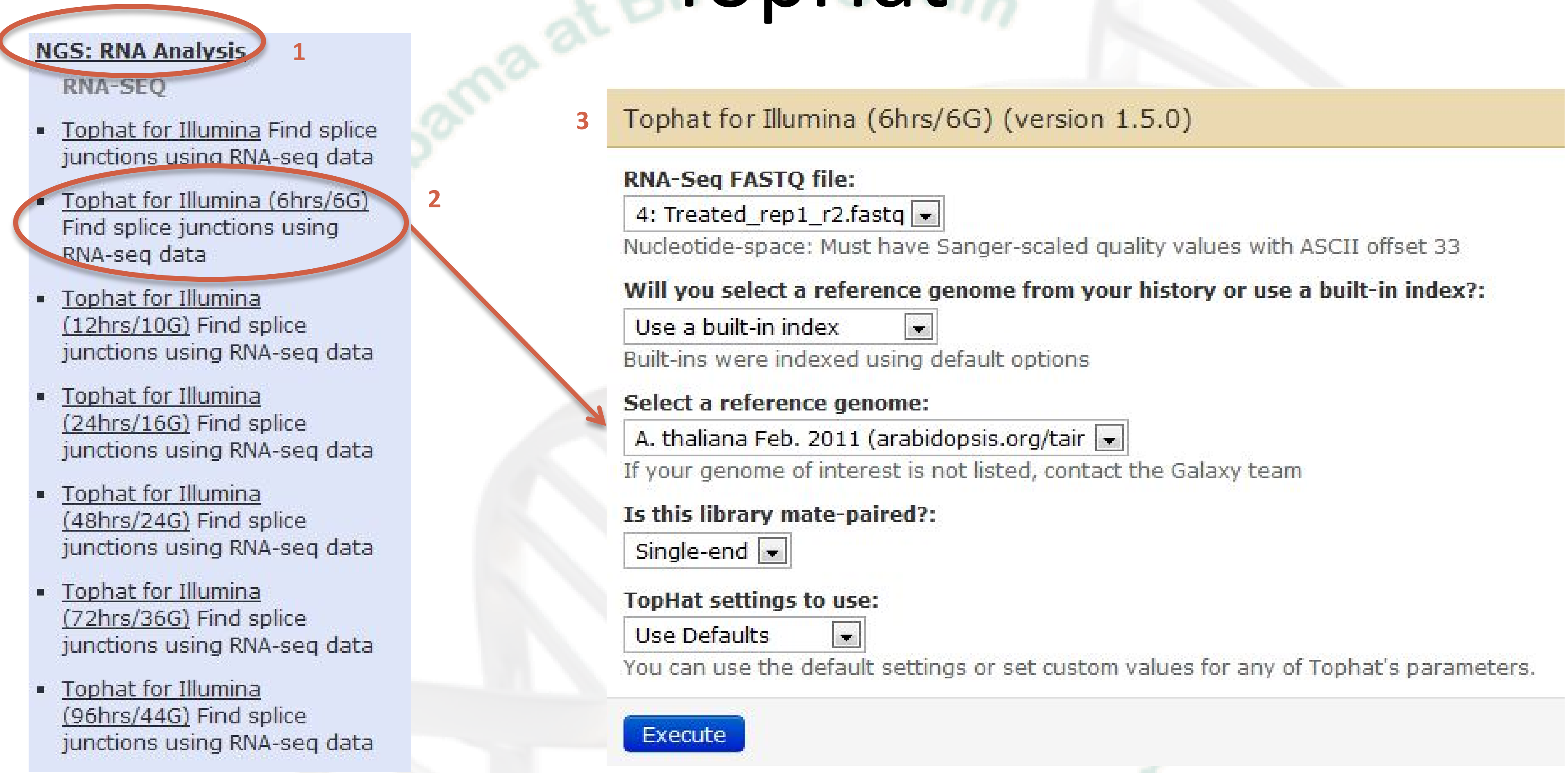
- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per sequence quality scores
- ✗ Per base sequence content
- ✗ Per base GC content
- ✗ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✗ Sequence Duplication Levels
- ✗ Overrepresented sequences
- ✗ Kmer Content

✓ 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



NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Tophat for Illumina (6hrs/6G) Find splice junctions using RNA-seq data
- Tophat for Illumina (12hrs/10G) Find splice junctions using RNA-seq data
- Tophat for Illumina (24hrs/16G) Find splice junctions using RNA-seq data
- Tophat for Illumina (48hrs/24G) Find splice junctions using RNA-seq data
- Tophat for Illumina (72hrs/36G) Find splice junctions using RNA-seq data
- Tophat for Illumina (96hrs/44G) Find splice junctions using RNA-seq data

Tophat for Illumina (6hrs/6G) (version 1.5.0)

RNA-Seq FASTQ file:
4: Treated_rep1_r2.fastq
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
Built-ins were indexed using default options

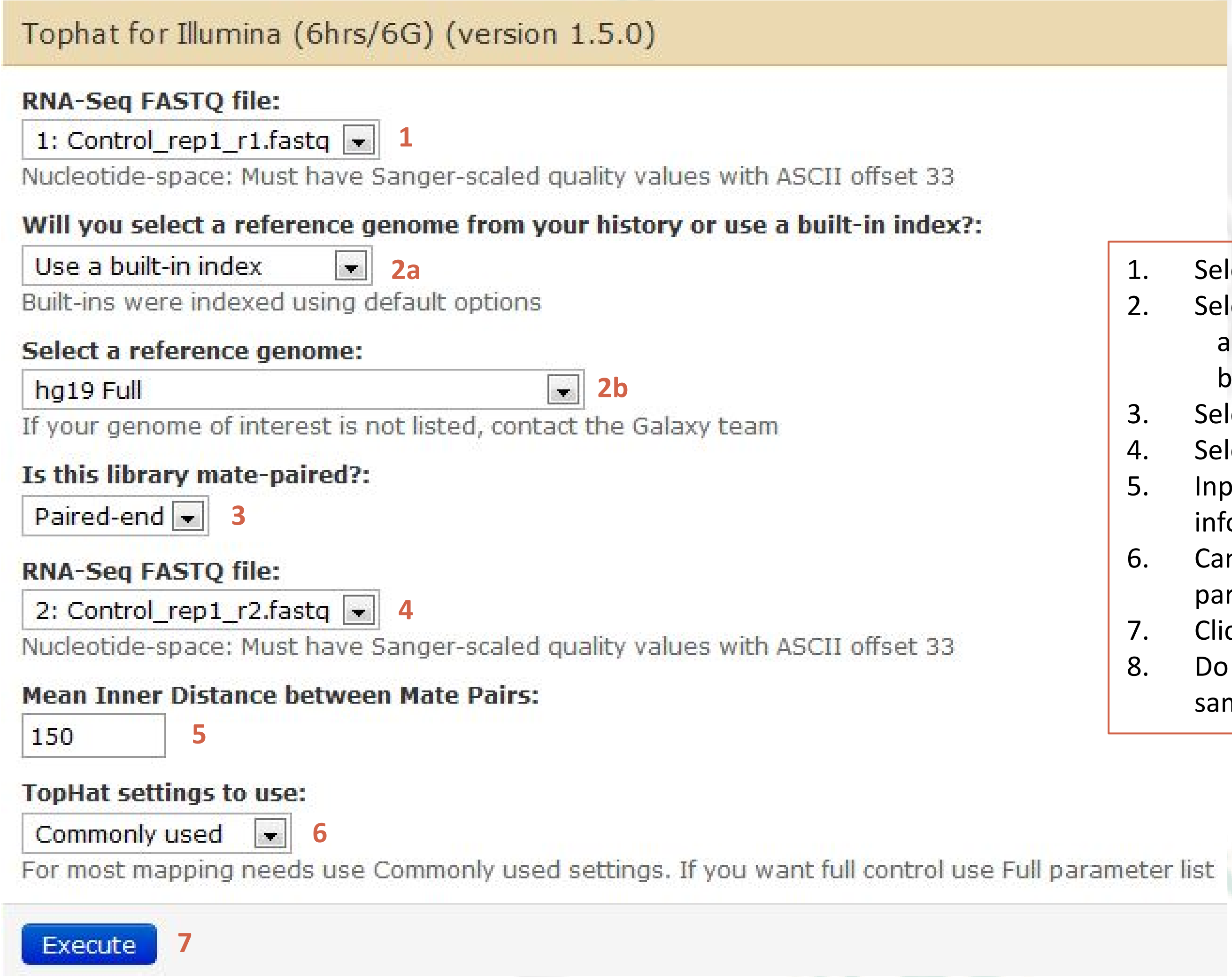
Select a reference genome:
A. thaliana Feb. 2011 (arabidopsis.org/tair
If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:
Single-end

TopHat settings to use:
Use Defaults
You can use the default settings or set custom values for any of Tophat's parameters.

1. Click on “NGS: RNA Analysis”
2. Click on “Tophat for Illumina (6hrs/6G)”
3. Default window with options appears

TopHat



Tophat for Illumina (6hrs/6G) (version 1.5.0)

RNA-Seq FASTQ file:
1: Control_rep1_r1.fastq
1
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Will you select a reference genome from your history or use a built-in index?:
Use a built-in index
2a
Built-ins were indexed using default options

Select a reference genome:
hg19 Full
2b
If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:
Paired-end
3

RNA-Seq FASTQ file:
2: Control_rep1_r2.fastq
4
Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs:
150
5

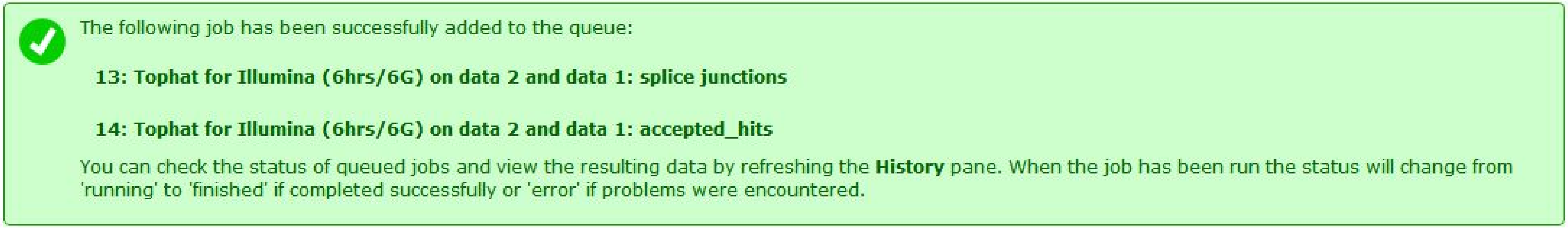
TopHat settings to use:
Commonly used
6
For most mapping needs use Commonly used settings. If you want full control use Full parameter list

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

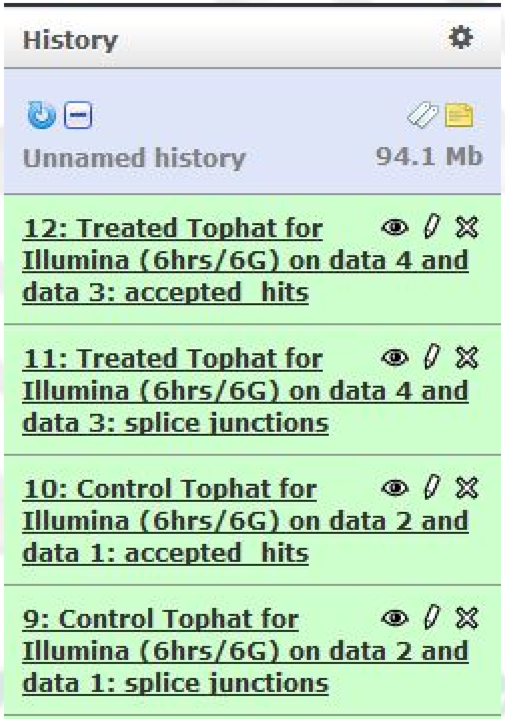


The following job has been successfully added to the queue:

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.



History

Unnamed history 94.1 Mb

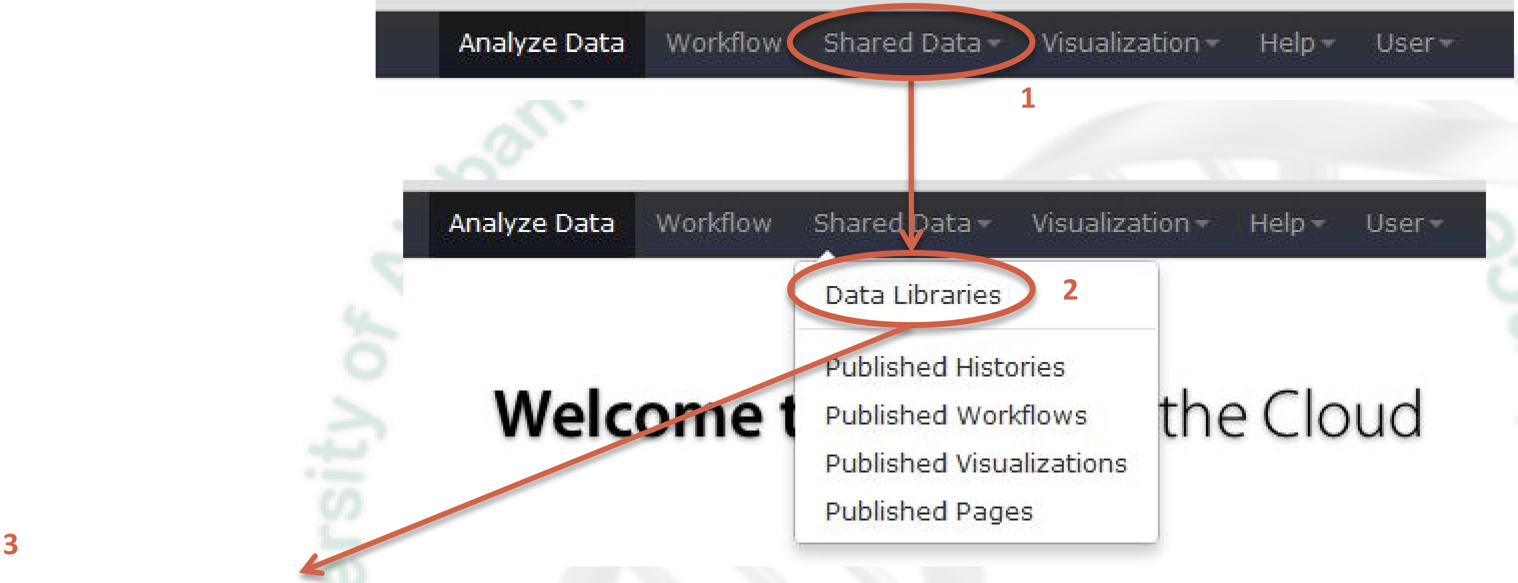
12: Treated Tophat for Illumina (6hrs/6G) on data 4 and data 3: accepted_hits

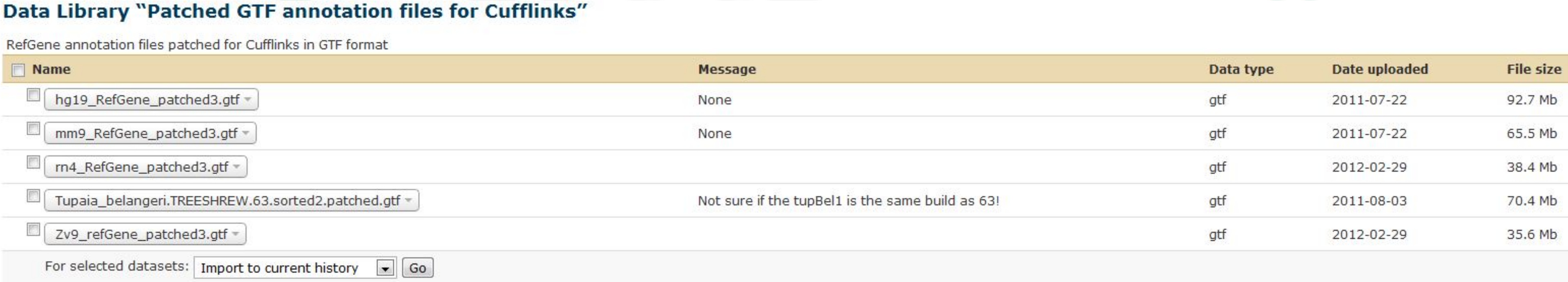
11: Treated Tophat for Illumina (6hrs/6G) on data 4 and data 3: splice junctions

10: Control Tophat for Illumina (6hrs/6G) on data 2 and data 1: accepted_hits

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

GTF Annotation Files



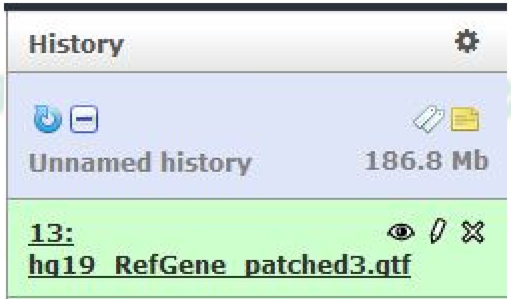


Data Library "Patched GTF annotation files for Cufflinks"

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

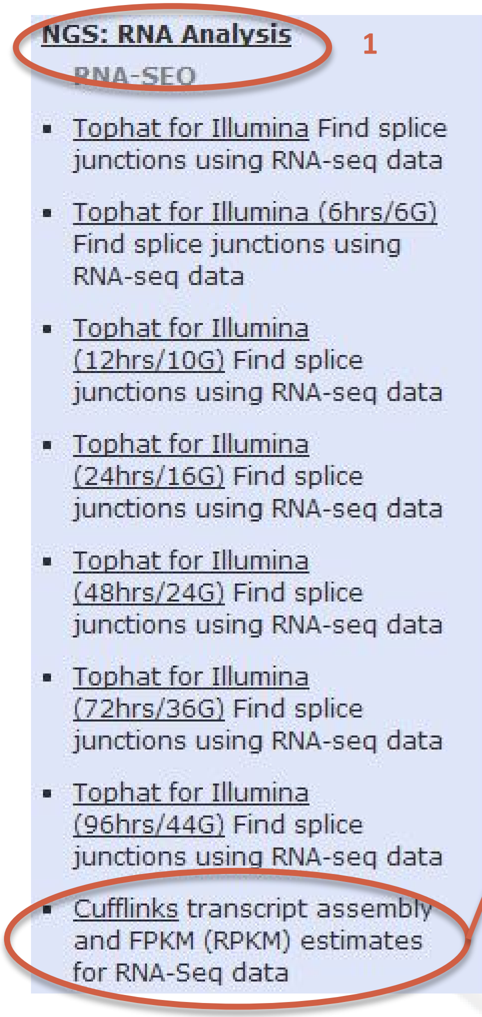


History

Unnamed history 186.8 Mb

13: hg19_RefGene_patched3.gtf

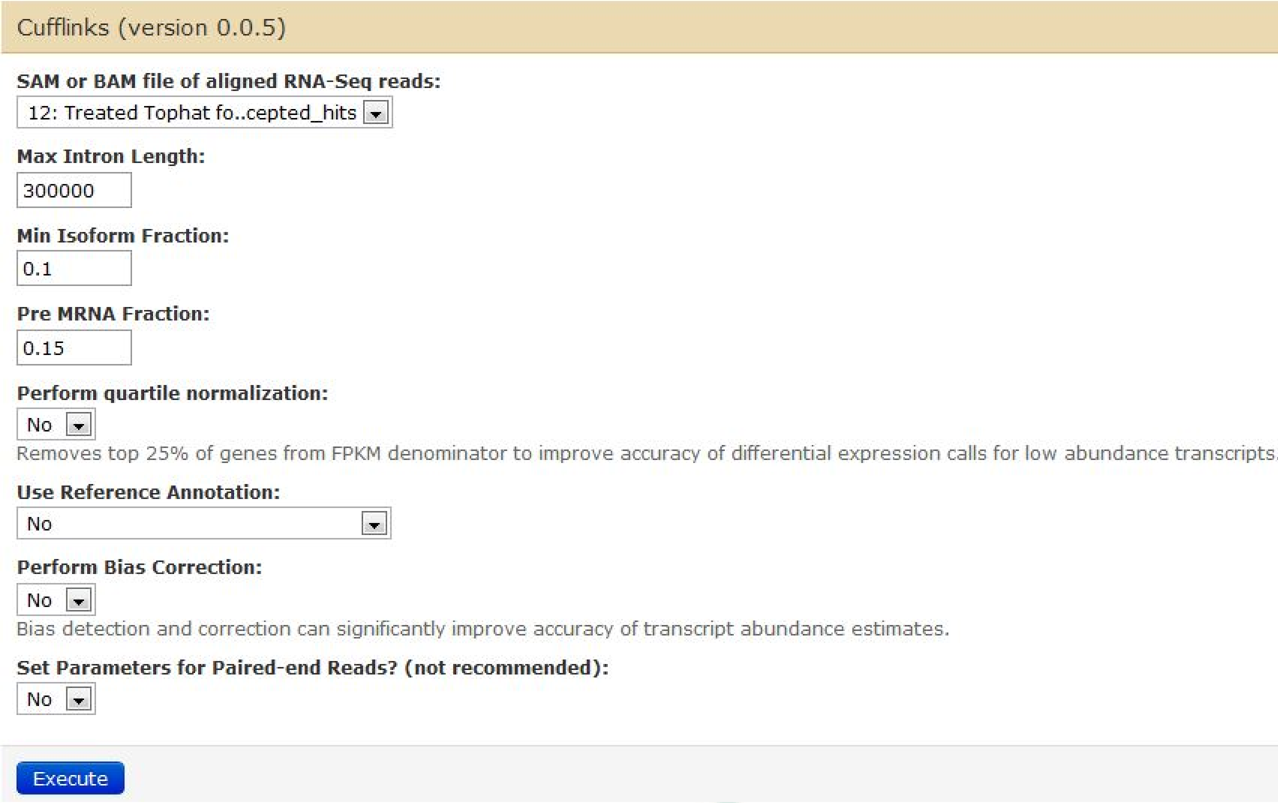
Cufflinks



NGS: RNA Analysis

RNA-SEQ

- Tophat for Illumina Find splice junctions using RNA-seq data
- Tophat for Illumina (6hrs/6G) Find splice junctions using RNA-seq data
- Tophat for Illumina (12hrs/10G) Find splice junctions using RNA-seq data
- Tophat for Illumina (24hrs/16G) Find splice junctions using RNA-seq data
- Tophat for Illumina (48hrs/24G) Find splice junctions using RNA-seq data
- Tophat for Illumina (72hrs/36G) Find splice junctions using RNA-seq data
- Tophat for Illumina (96hrs/44G) Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data



Cufflinks (version 0.0.5)

SAM or BAM file of aligned RNA-Seq reads: 12: Treated Tophat fo..cepted_hits

Max Intron Length: 300000

Min Isoform Fraction: 0.1

Pre MRNA Fraction: 0.15

Perform quartile normalization: No

Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Use Reference Annotation: No

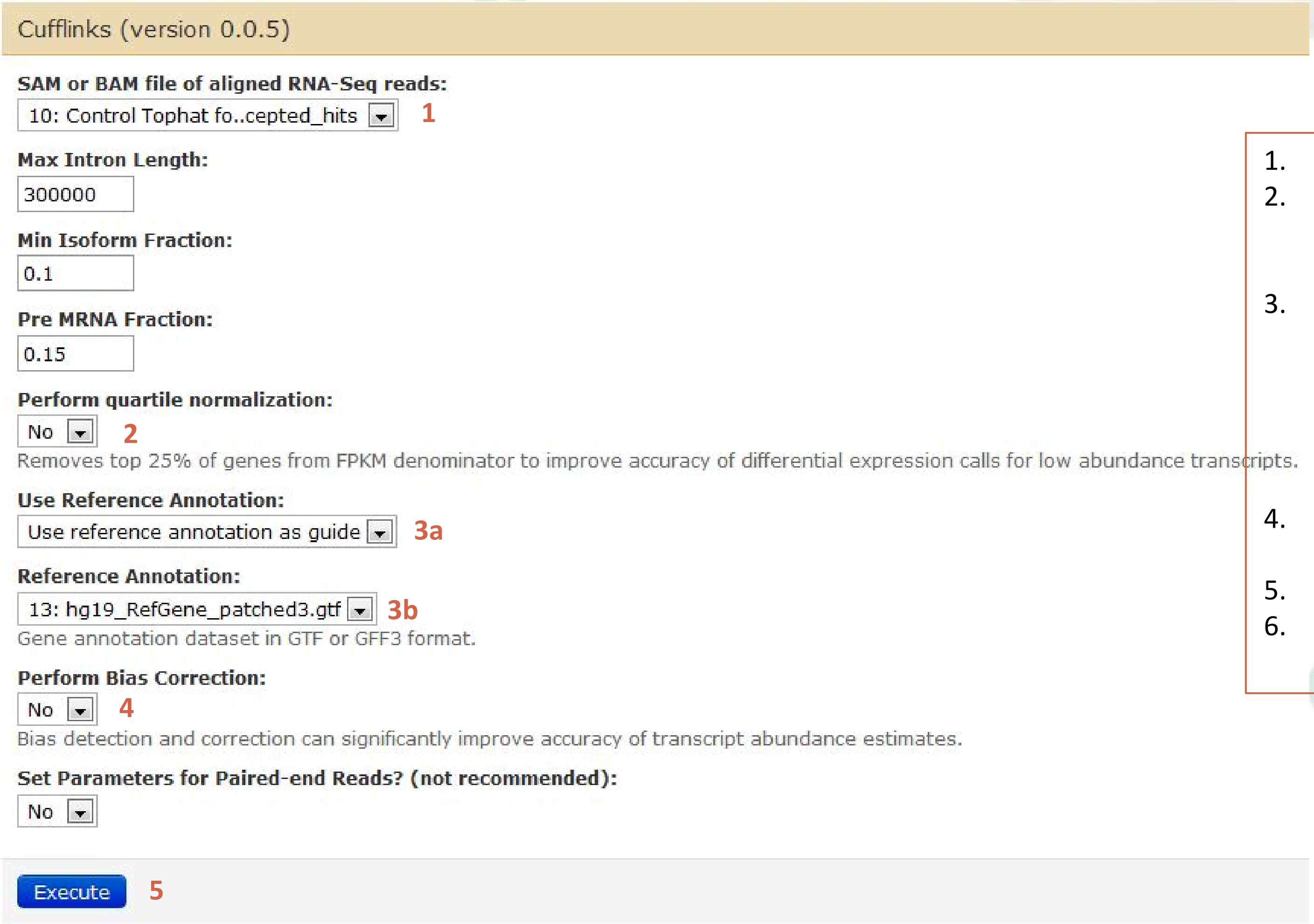
Perform Bias Correction: No

Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Set Parameters for Paired-end Reads? (not recommended): No

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

Cufflinks



Cufflinks (version 0.0.5)

SAM or BAM file of aligned RNA-Seq reads:
10: Control Tophat fo..cepted_hits ▾ 1

Max Intron Length:
300000

Min Isoform Fraction:
0.1

Pre MRNA Fraction:
0.15

Perform quartile normalization:
No ▾ 2
Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Use Reference Annotation:
Use reference annotation as guide ▾ 3a

Reference Annotation:
13: hg19_RefGene_patched3.gtf ▾ 3b
Gene annotation dataset in GTF or GFF3 format.

Perform Bias Correction:
No ▾ 4
Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Set Parameters for Paired-end Reads? (not recommended):
No ▾

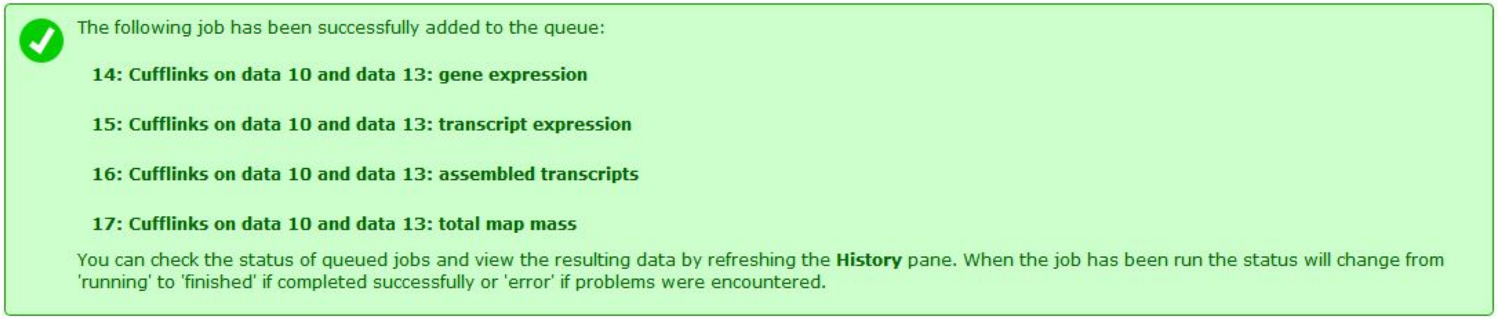
5

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...

Cufflinks output files



The following job has been successfully added to the queue:

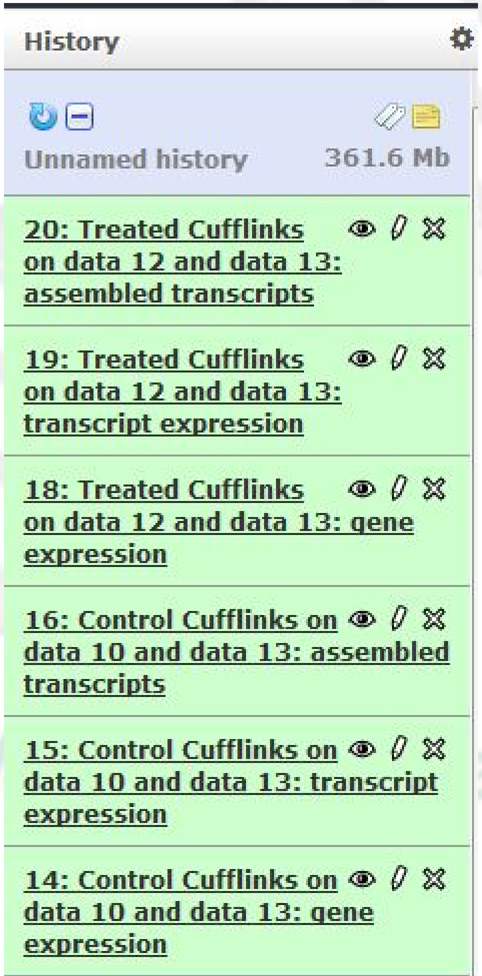
14: Cufflinks on data 10 and data 13: gene expression

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.



Unnamed history 361.6 Mb

20: Treated Cufflinks on data 12 and data 13: assembled transcripts

19: Treated Cufflinks on data 12 and data 13: transcript expression

18: Treated Cufflinks on data 12 and data 13: gene expression

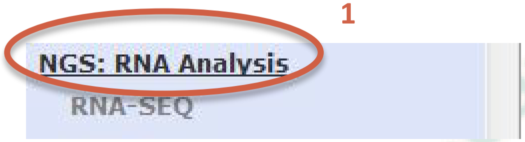
16: Control Cufflinks on data 10 and data 13: assembled transcripts

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

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

Cuffmerge

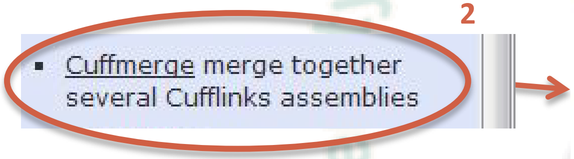
1



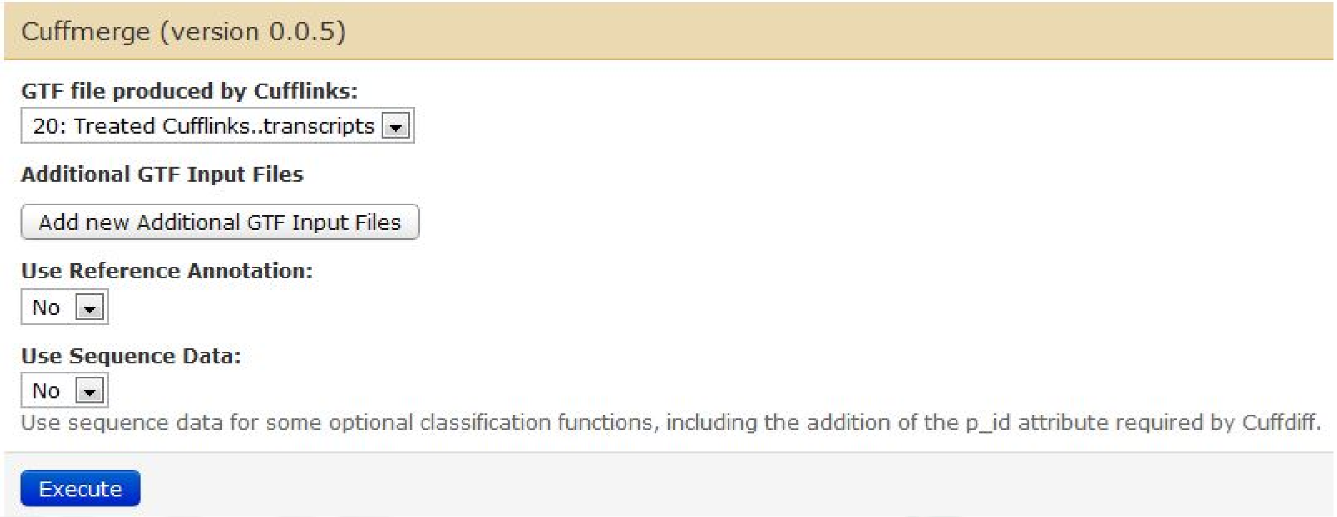
NGS: RNA Analysis

RNA-SEQ

2



- Cuffmerge merge together several Cufflinks assemblies



Cuffmerge (version 0.0.5)

GTF file produced by Cufflinks: 20: Treated Cufflinks..transcripts

Additional GTF Input Files

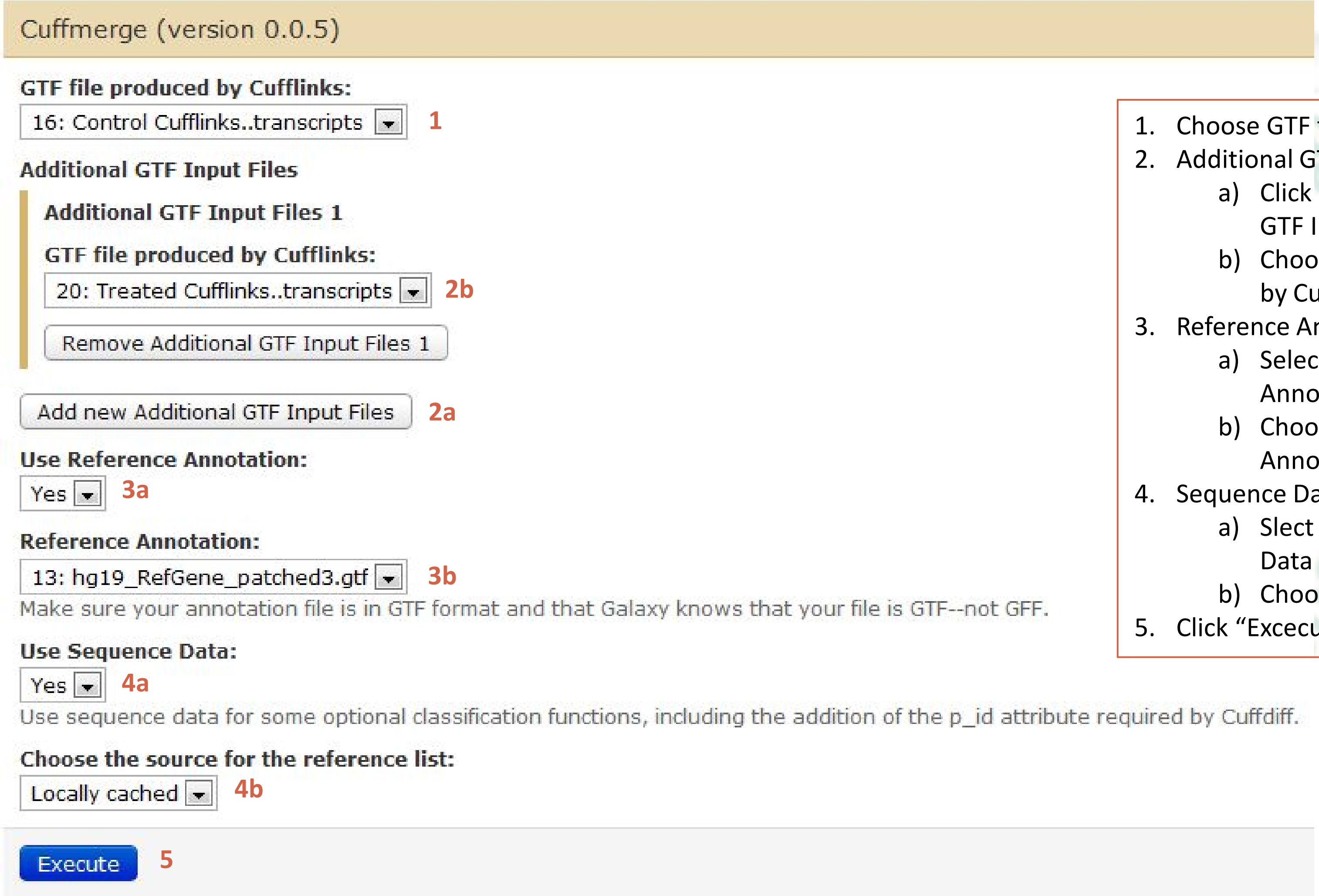
Use Reference Annotation: No

Use Sequence Data: No

Use sequence data for some optional classification functions, including the addition of the p_id attribute required by Cuffdiff.

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

Cuffmerge



Cuffmerge (version 0.0.5)

GTF file produced by Cufflinks:

16: Control Cufflinks..transcripts

Additional GTF Input Files

Additional GTF Input Files 1

GTF file produced by Cufflinks:

20: Treated Cufflinks..transcripts

Remove Additional GTF Input Files 1

Add new Additional GTF Input Files

Use Reference Annotation:

Yes

Reference Annotation:

13: hg19_RefGene_patched3.gtf

Make sure your annotation file is in GTF format and that Galaxy knows that your file is GTF--not GFF.

Use Sequence Data:

Yes

Use sequence data for some optional classification functions, including the addition of the p_id attribute required by Cuffdiff.

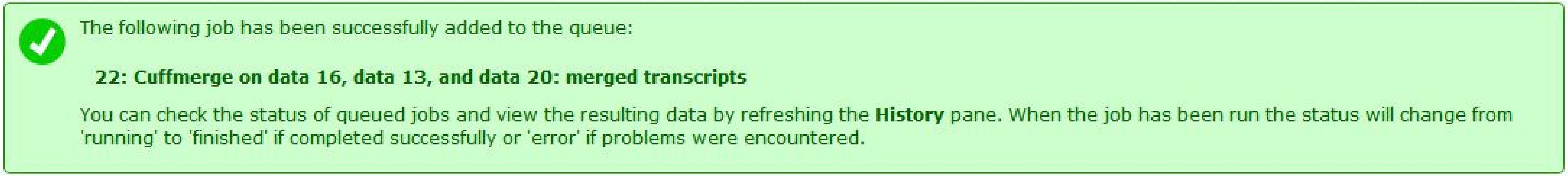
Choose the source for the reference list:

Locally cached

Execute

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”

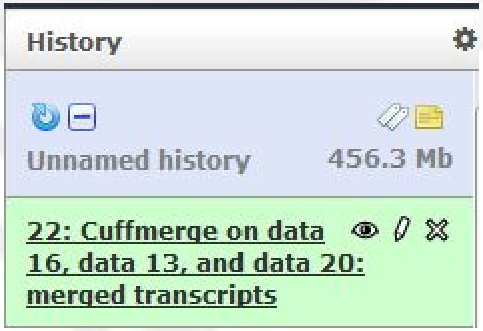
Cuffmerge output files



The following job has been successfully added to the queue:

22: Cuffmerge on data 16, data 13, and data 20: merged transcripts

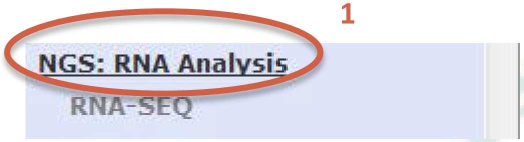
You can check the status of queued jobs and view the resulting data by refreshing the **History** pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.



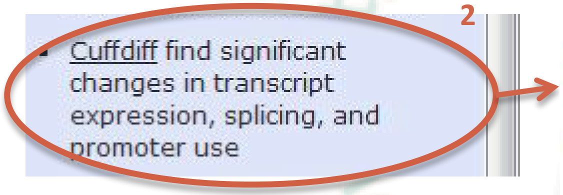
Unnamed history 456.3 Mb

22: Cuffmerge on data 16, data 13, and data 20: merged transcripts

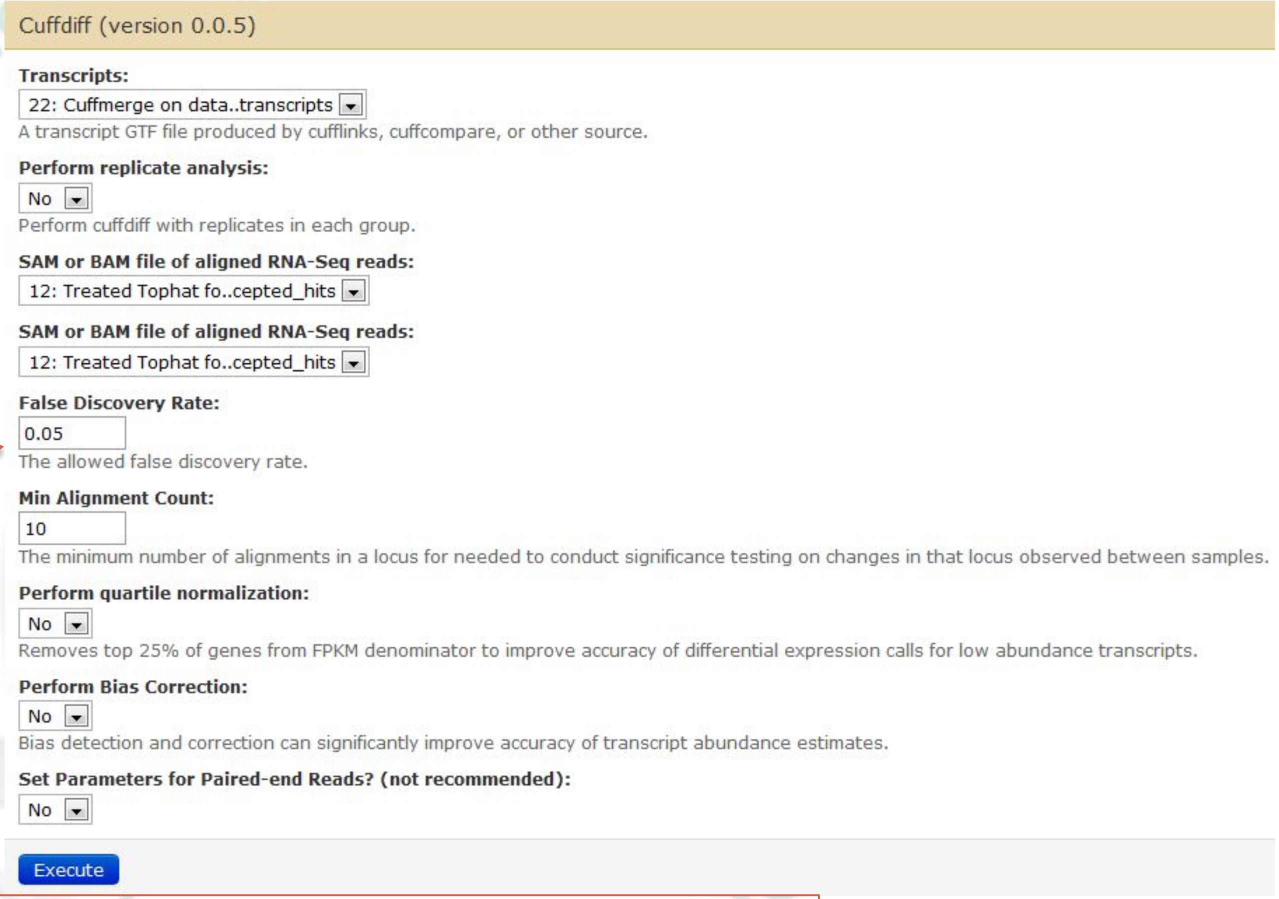
Cuffdiff



NGS: RNA Analysis
RNA-SEQ



- Cuffdiff find significant changes in transcript expression, splicing, and promoter use



Cuffdiff (version 0.0.5)

Transcripts:
22: Cuffmerge on data..transcripts
A transcript GTF file produced by cufflinks, cuffcompare, or other source.

Perform replicate analysis:
No
Perform cuffdiff with replicates in each group.

SAM or BAM file of aligned RNA-Seq reads:
12: Treated Tophat fo..cepted_hits

SAM or BAM file of aligned RNA-Seq reads:
12: Treated Tophat fo..cepted_hits

False Discovery Rate:
0.05
The allowed false discovery rate.

Min Alignment Count:
10
The minimum number of alignments in a locus for needed to conduct significance testing on changes in that locus observed between samples.

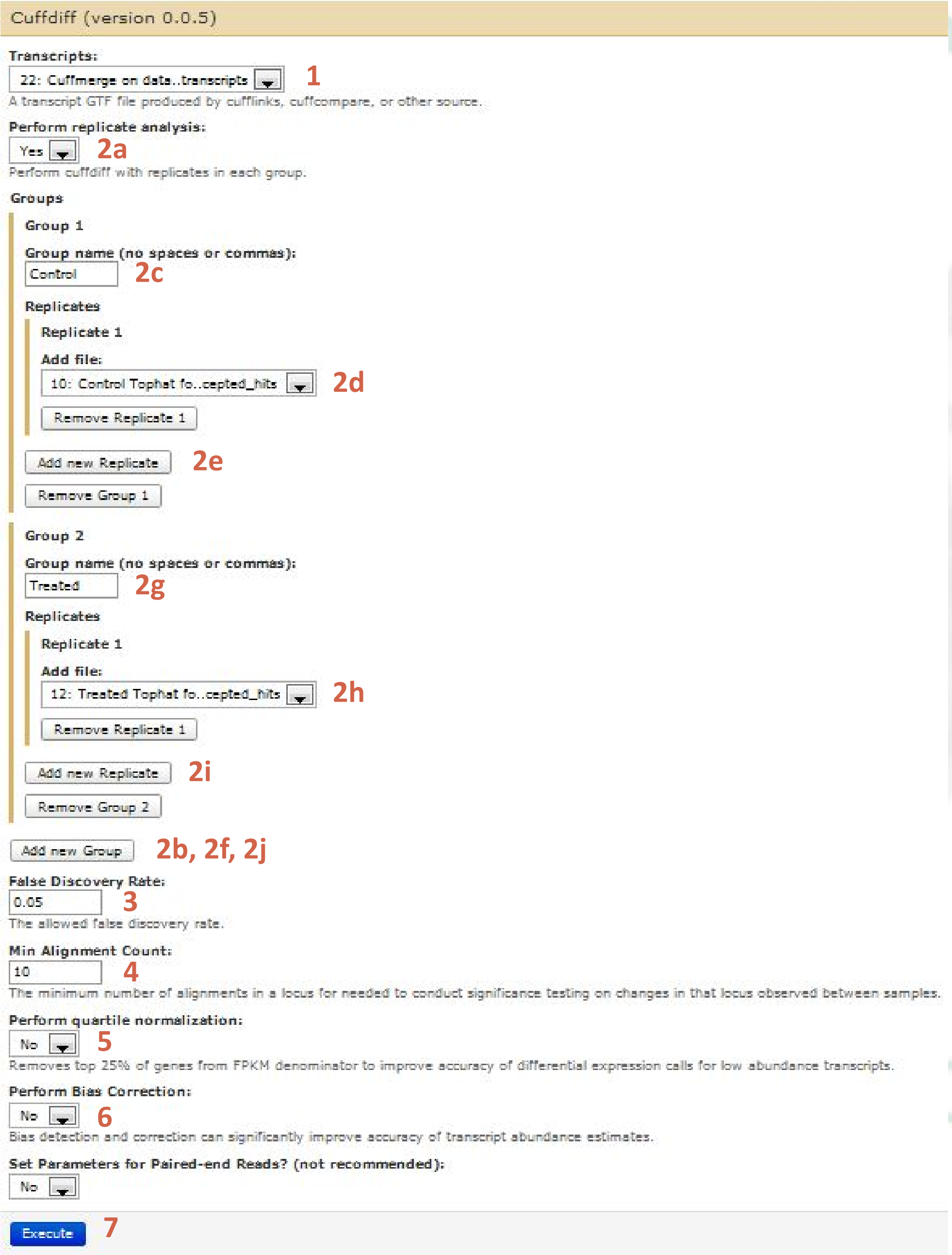
Perform quartile normalization:
No
Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Perform Bias Correction:
No
Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Set Parameters for Paired-end Reads? (not recommended):
No

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

Cuffdiff



Cuffdiff (version 0.0.5)

Transcripts:
22: Cuffmerge on data..transcripts
A transcript GTF file produced by cufflinks, cuffcompare, or other source.

Perform replicate analysis:
Yes
Perform cuffdiff with replicates in each group.

Groups

Group 1
Group name (no spaces or commas):
Control
Replicates
Replicate 1
Add file:
10: Control Tophat fo..cepted_hits

Group 2
Group name (no spaces or commas):
Treated
Replicates
Replicate 1
Add file:
12: Treated Tophat fo..cepted_hits

False Discovery Rate:
0.05
The allowed false discovery rate.

Min Alignment Count:
10
The minimum number of alignments in a locus for needed to conduct significance testing on changes in that locus observed between samples.

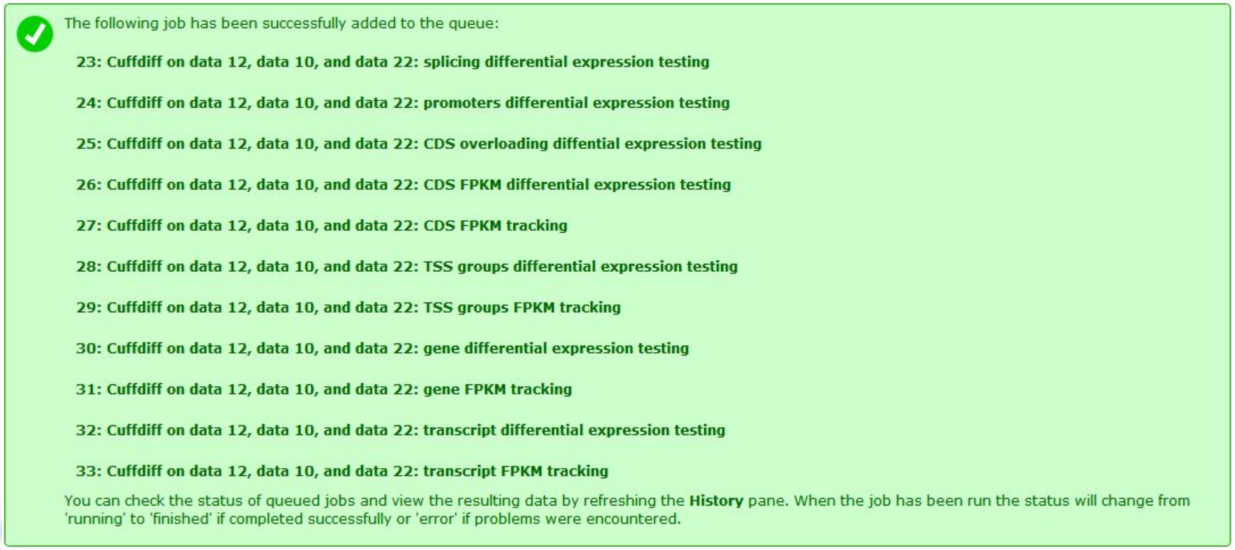
Perform quartile normalization:
No
Removes top 25% of genes from FPKM denominator to improve accuracy of differential expression calls for low abundance transcripts.

Perform Bias Correction:
No
Bias detection and correction can significantly improve accuracy of transcript abundance estimates.

Set Parameters for Paired-end Reads? (not recommended):
No

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

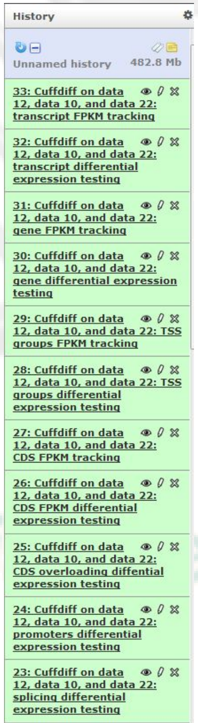
30: Cuffdiff on data 12, data 10, and data 22: gene differential expression testing

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.



History

Unnamed history 482.8 Mb

33: Cuffdiff on data 12, data 10, and data 22: transcript FPKM tracking

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

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

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

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

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

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

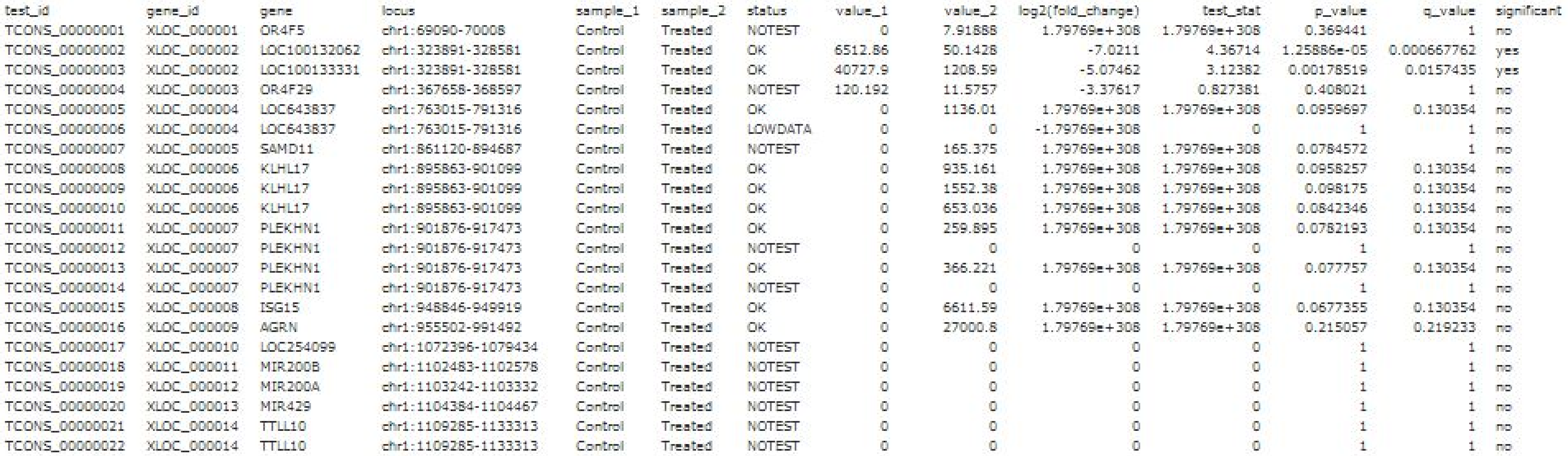
26: Cuffdiff on data 12, data 10, and data 22: CDS FPKM differential expression testing

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

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

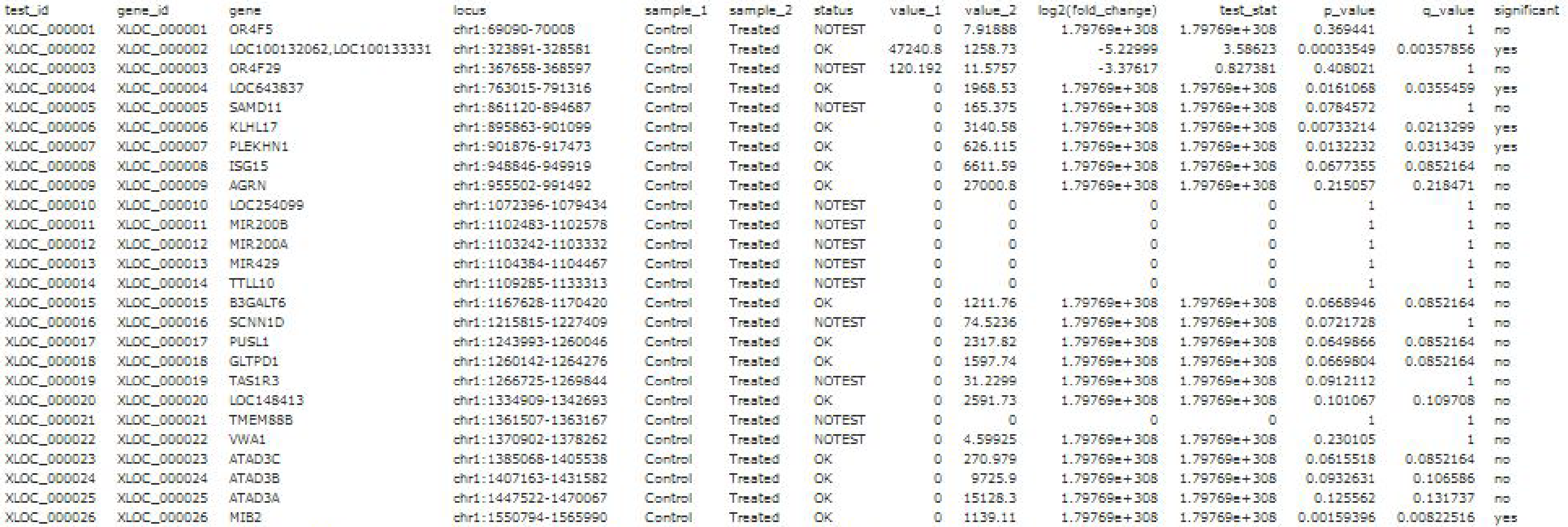
23: Cuffdiff on data 12, data 10, and data 22: splicing differential expression testing

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

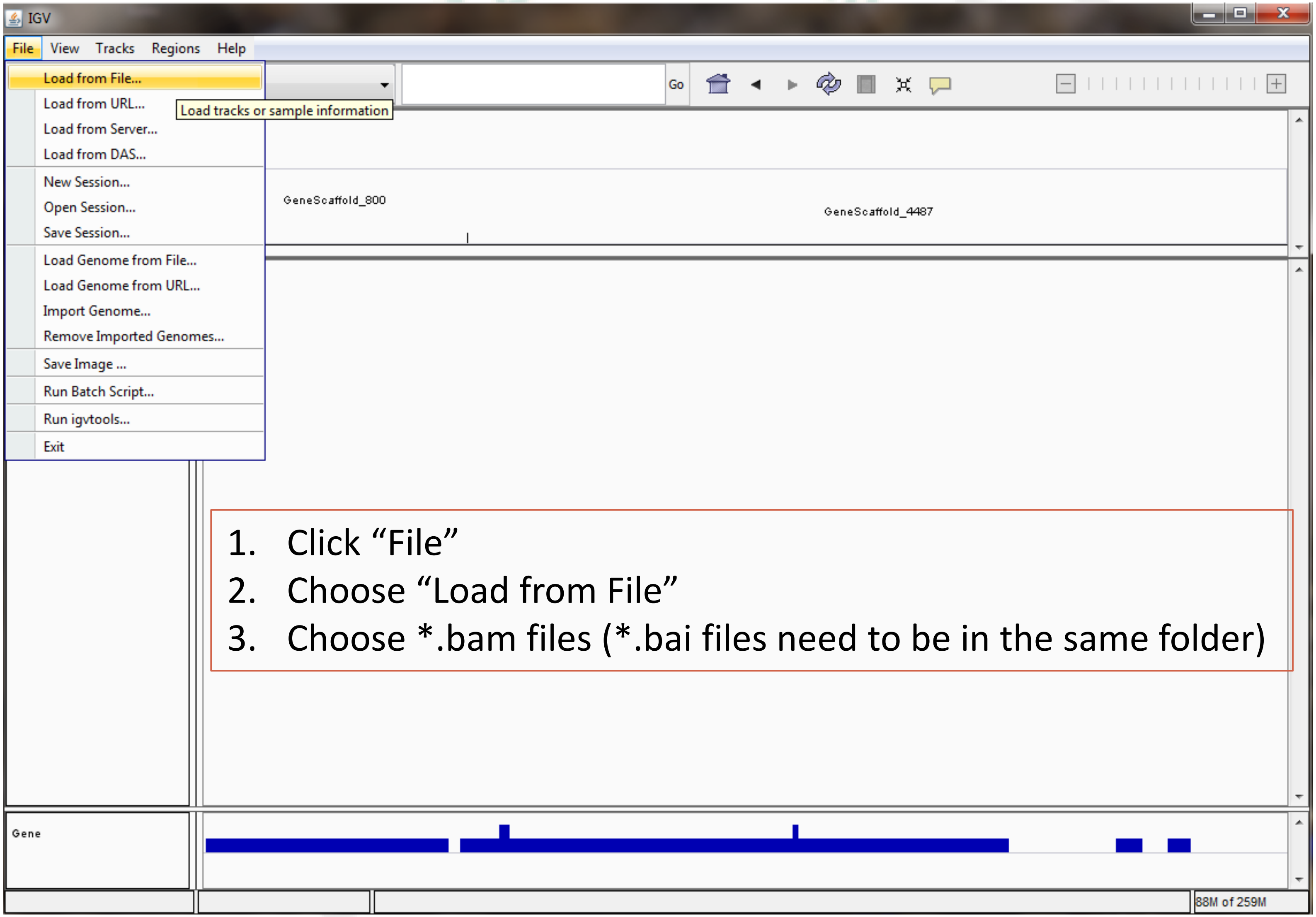


test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
XLOC_000001	XLOC_000001	OR4F5	chr1:69090-70008	Control	Treated	NOTEST	0	7.91888	1.79769e+308	1.79769e+308	0.369441	1	no
XLOC_000002	XLOC_000002	LOC100132062,LOC100133331	chr1:323891-328581	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

Contents

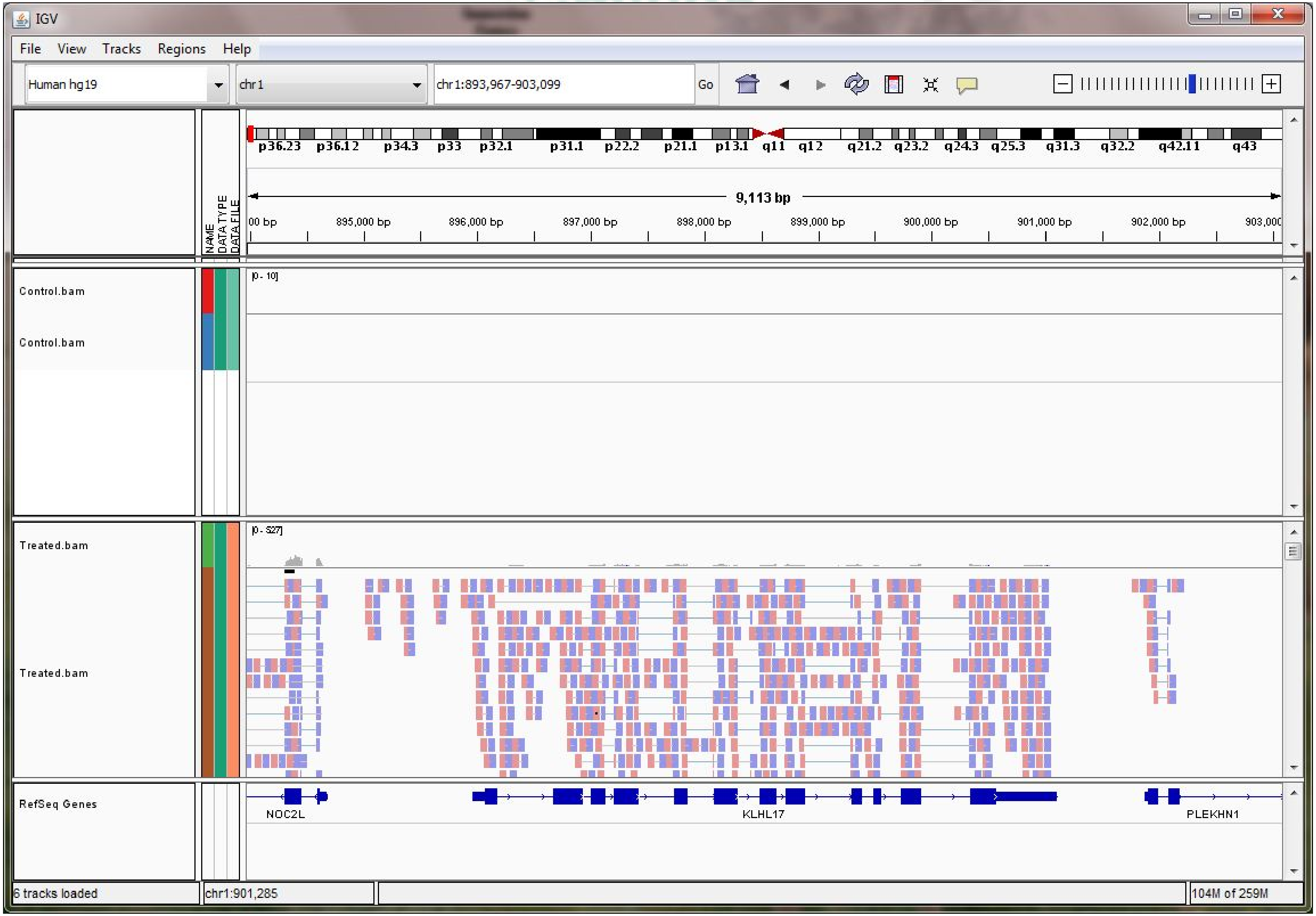
- FASTQ anatomy
- RNA-Seq demo
- Genomics Viewer (IGV) demo
- Whole Genome/Exome demo
- References and web links

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)

IGV



Human hg19

chr1

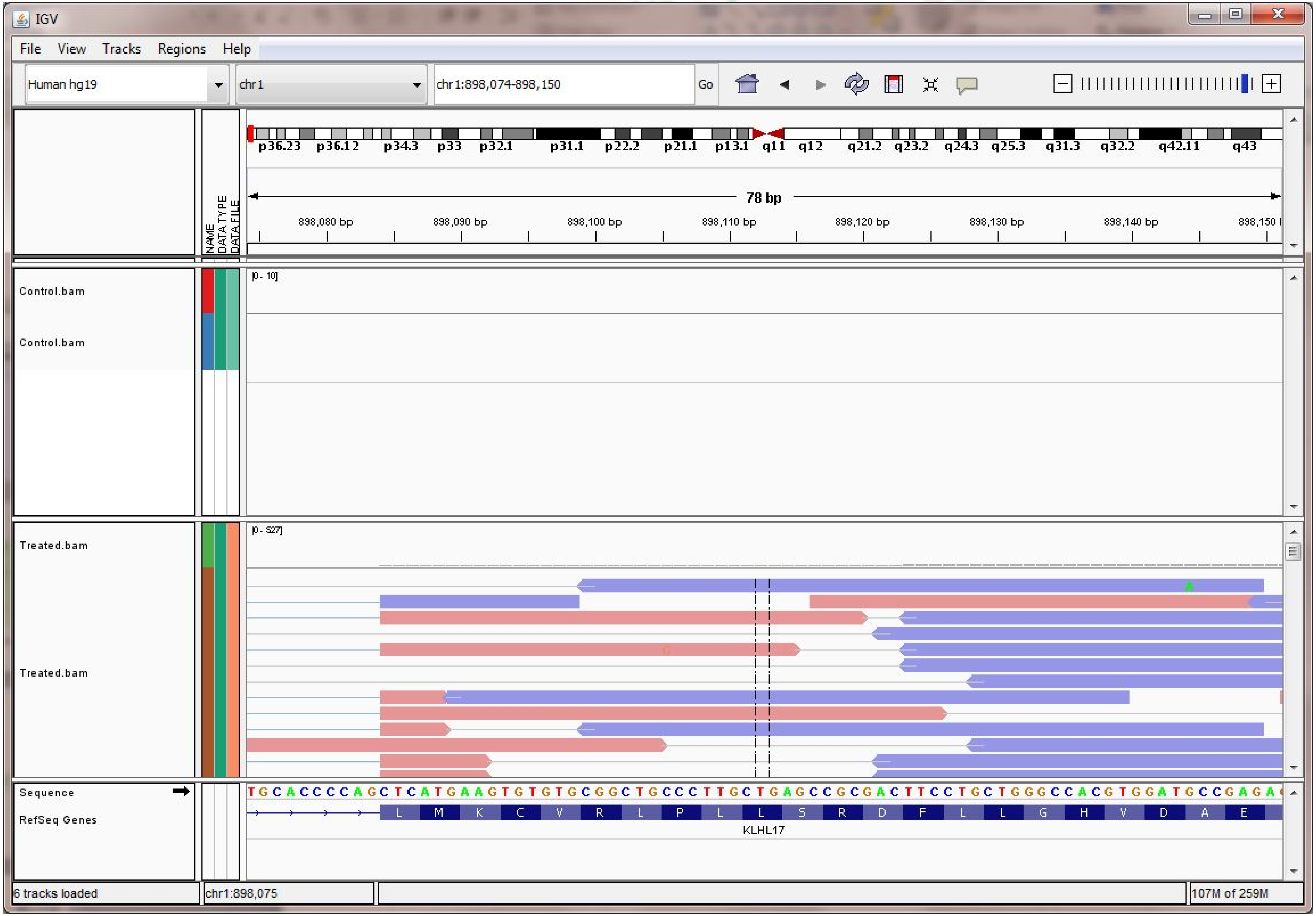
chr1:893,967-903,099

6 tracks loaded

chr1:901,285

104M of 259M

IGV



Human hg19

chr1

chr1:898,074-898,150

6 tracks loaded

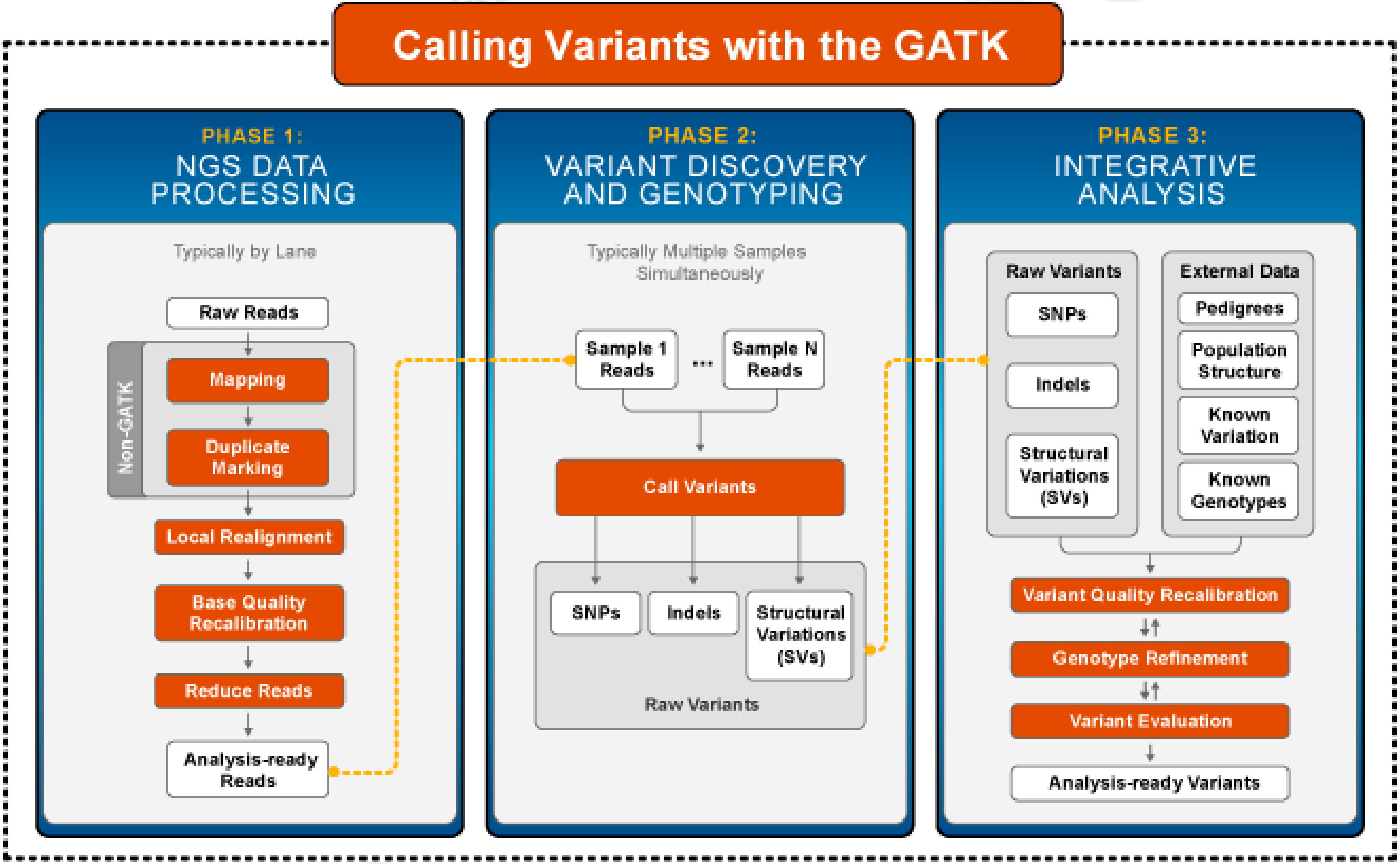
chr1:898,075

107M of 259M

Contents

- FASTQ anatomy
- RNA-Seq demo
- Genomics Viewer (IGV) demo
- Whole Genome/Exome demo
- References and web links

Whole Genome/Exome GATK pipeline



Calling Variants with the GATK

PHASE 1: NGS DATA PROCESSING

Typically by Lane

Raw Reads

Non-GATK

Mapping

Duplicate Marking

Local Realignment

Base Quality Recalibration

Reduce Reads

Analysis-ready Reads

PHASE 2: VARIANT DISCOVERY AND GENOTYPING

Typically Multiple Samples Simultaneously

Sample 1 Reads ... Sample N Reads

Call Variants

SNPs

Indels

Structural Variations (SVs)

Raw Variants

PHASE 3: INTEGRATIVE ANALYSIS

Raw Variants

SNPs

Indels

Structural Variations (SVs)

External Data

Pedigrees

Population Structure

Known Variation

Known Genotypes

Variant Quality Recalibration

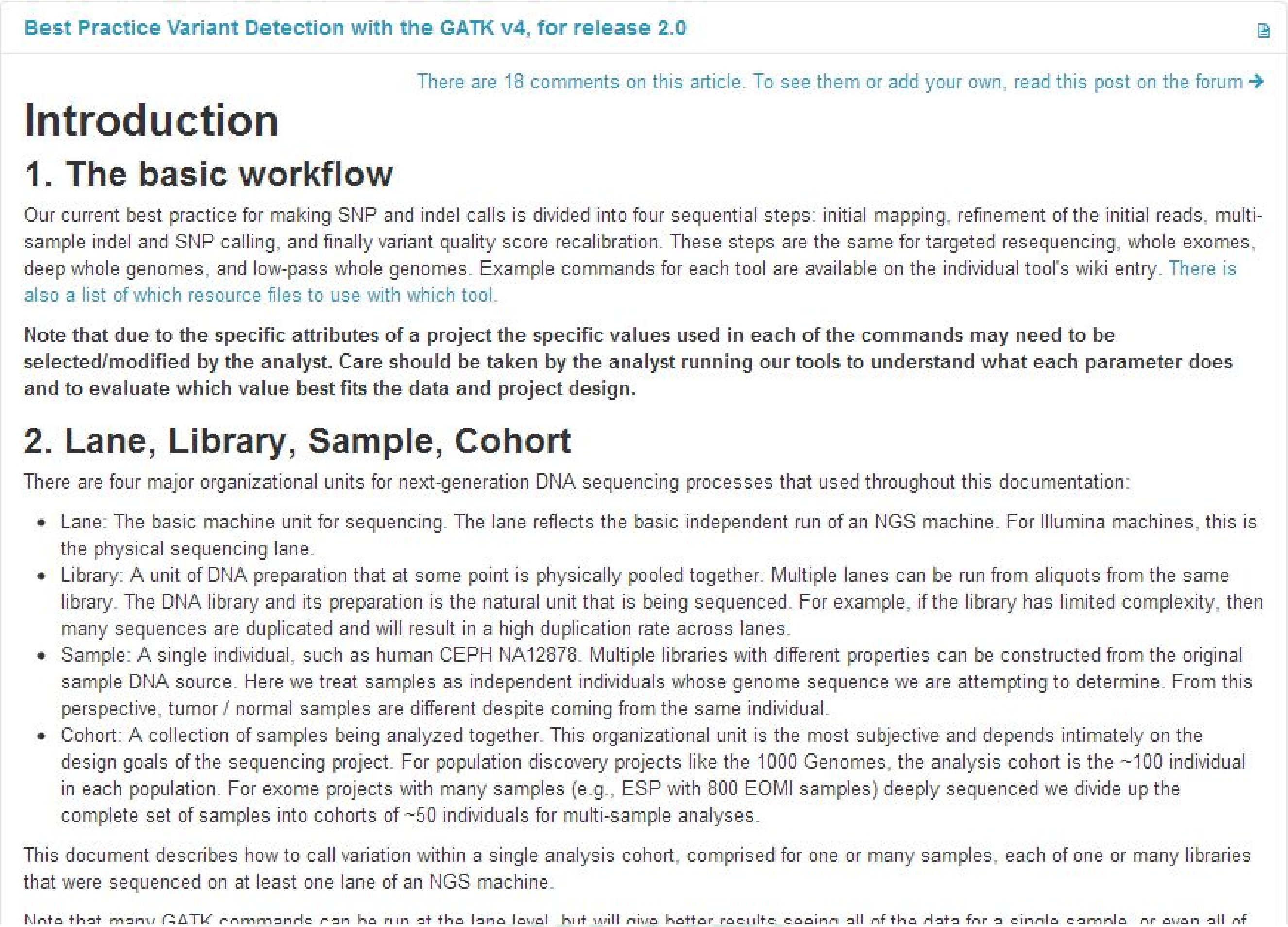
Genotype Refinement

Variant Evaluation

Analysis-ready Variants

GATK Best Practices

(http://www.broadinstitute.org/gatk/)



Best Practice Variant Detection with the GATK v4, for release 2.0

There are 18 comments on this article. To see them or add your own, read this post on the forum →

Introduction

1. The basic workflow

Our current best practice for making SNP and indel calls is divided into four sequential steps: initial mapping, refinement of the initial reads, multi-sample indel and SNP calling, and finally variant quality score recalibration. These steps are the same for targeted resequencing, whole exomes, deep whole genomes, and low-pass whole genomes. Example commands for each tool are available on the individual tool's wiki entry. There is also a list of which resource files to use with which tool.

Note that due to the specific attributes of a project the specific values used in each of the commands may need to be selected/modified by the analyst. Care should be taken by the analyst running our tools to understand what each parameter does and to evaluate which value best fits the data and project design.

2. Lane, Library, Sample, Cohort

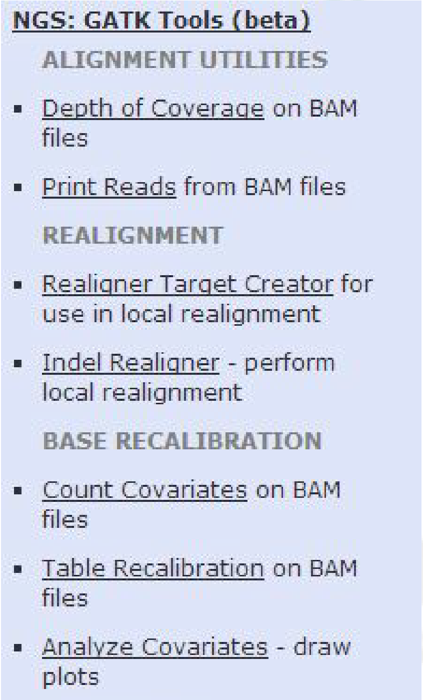
There are four major organizational units for next-generation DNA sequencing processes that used throughout this documentation:

- Lane: The basic machine unit for sequencing. The lane reflects the basic independent run of an NGS machine. For Illumina machines, this is the physical sequencing lane.
- Library: A unit of DNA preparation that at some point is physically pooled together. Multiple lanes can be run from aliquots from the same library. The DNA library and its preparation is the natural unit that is being sequenced. For example, if the library has limited complexity, then many sequences are duplicated and will result in a high duplication rate across lanes.
- Sample: A single individual, such as human CEPH NA12878. Multiple libraries with different properties can be constructed from the original sample DNA source. Here we treat samples as independent individuals whose genome sequence we are attempting to determine. From this perspective, tumor / normal samples are different despite coming from the same individual.
- Cohort: A collection of samples being analyzed together. This organizational unit is the most subjective and depends intimately on the design goals of the sequencing project. For population discovery projects like the 1000 Genomes, the analysis cohort is the ~100 individual in each population. For exome projects with many samples (e.g., ESP with 800 EOMI samples) deeply sequenced we divide up the complete set of samples into cohorts of ~50 individuals for multi-sample analyses.

This document describes how to call variation within a single analysis cohort, comprised for one or many samples, each of one or many libraries that were sequenced on at least one lane of an NGS machine.

Note that many GATK commands can be run at the lane level, but will give better results seeing all of the data for a single sample, or even all of

GATK (beta) on Galaxy



NGS: GATK Tools (beta)

ALIGNMENT UTILITIES

- Depth of Coverage on BAM files

6 ▪ Print Reads from BAM files

REALIGNMENT

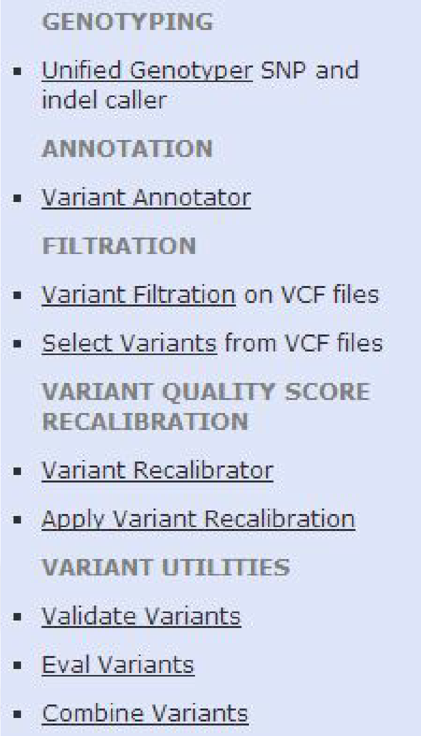
3 ▪ Realigner Target Creator for use in local realignment

4 ▪ Indel Realigner - perform local realignment

BASE RECALIBRATION

5 ▪ Count Covariates on BAM files

- Table Recalibration on BAM files
- Analyze Covariates - draw plots



GENOTYPING

7 ▪ Unified Genotyper SNP and indel caller

ANNOTATION

- Variant Annotator

FILTRATION

- Variant Filtration on VCF files

11 ▪ Select Variants from VCF files

VARIANT QUALITY SCORE RECALIBRATION

8 ▪ Variant Recalibrator

9 ▪ Apply Variant Recalibration

VARIANT UTILITIES

- Validate Variants
- Eval Variants

10 ▪ Combine Variants

Basic Steps* (options are up to you):

1. BWA alignment
2. Mark duplicates (Picard)
3. Realigner Target Creator
4. Indel Realigner
5. Base Recalibrator (Count Covariates)
6. Print Reads
7. Unified Genotyper (new in Ver2 is Haplotype Caller) (SNPs and Indels done separately)
8. Variant Recalibrator (SNPs and Indels done separately)
9. Apply Recalibration (SNPs and Indels done separately)
10. Combine Variants
11. Select Variants
12. Compare/contrast variants
13. snpEFF

* This follows the **basic** pipeline shown 2 slides ago. Each project is different and may need additional tools to answer the biological question(s). Also, options for each tool will vary as well.

Contents

- FASTQ anatomy
- RNA-Seq demo
- Genomics Viewer (IGV) demo
- Whole Genome/Exome demo
- References and web links

References and web links

- TopHat
  - Trapnell C, Pachter L, Salzberg SL. **TopHat: discovering splice junctions with RNA-Seq**.*Bioinformatics* doi:10.1093/bioinformatics/btp120
  - http://tophat.cbcb.umd.edu/
- Bowtie
  - Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10:R25.
  - http://bowtie-bio.sourceforge.net/index.shtml
- Cufflinks
  - Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation** *Nature Biotechnology* doi:10.1038/nbt.1621
  - Roberts A, Trapnell C, Donaghey J, Rinn JL, Pachter L. **Improving RNA-Seq expression estimates by correcting for fragment bias** *Genome Biology* doi:10.1186/gb-2011-12-3-r22
  - Roberts A, Pimentel H, Trapnell C, Pachter L.**Identification of novel transcripts in annotated genomes using RNA-Seq** *Bioinformatics* doi:10.1093/bioinformatics/btr355
  - http://cufflinks.cbcb.umd.edu/
- TopHat and Cufflinks protocol
  - Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. **Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks** Nature Protocols **7**, 562-578 (2012) doi:10.1038/nprot.2012.016
- IGV
  - http://www.broadinstitute.org/igv/

Thanks! Questions?

Contact info:

David K. Crossman, Ph.D.
Bioinformatics Director
Heflin Center for Genomic Science
University of Alabama at Birmingham
http://www.heflingenetics.uab.edu
dkcrossm@uab.edu