Analysis of next generation sequencing experiments with Galaxy

March 24, 2011

Previous Hot Topics on Next Generation Sequencing Analysis

- Mapping next generation sequence reads

  http://iona.wi.mit.edu/bio/education/hot_topics/shortRead_mapping/Mapping_HTseq.pdf
- Analysis of ChIP-seq experiments

  http://iona.wi.mit.edu/bio/education/hot_topics/ChIPseq/ChIPSeq_HotTopics.pdf
- RNA-seq: Methods and Applications

  http://iona.wi.mit.edu/bio/education/hot_topics/RNAseq/RNA_Seq.pdf

Talk Outline

- Introduction to Galaxy
- Data upload
- Format conversion and quality control tools
- Analysis of ChIP-seq experiments with MACS
- Analysis of RNA-seq experiments with Tuxedo tools
- Demo

What is Galaxy

- A web based platform for analysis of large genomic datasets
- No need of programming experience.
- Integrates many tools within one interface:
  - Easy retrieval of data from UCSC, Biomart and other DBs
  - Powerful text manipulation tools (data preparation)
  - Filter on columns, join, sort, compute etc
  - Format conversion tools (text, tab, bed, GFF …)
  - Integrates tools from other sources. Ex: EMBOSS
  - MSA tools
  - Visualize data in UCSC browser.

    (See Hot topics Dec 09, http://iona.wi.mit.edu/bio/education/hot_topics/galaxy/Galaxy.pdf)

  - **Next Generation Sequencing Toolbox**

Documentation and Tutorials

- OpenHelix tutorials and exercises http://www.openhelix.com/cgi/tutorialInfo.cgi?id=82
- Galaxy tutorials http://galaxy.psu.edu/screencasts.html
- References

Galaxy developers: The Center for Comparative Genomics & Bioinformatics, Pennsylvania State University

Giardine, B., et al. Galaxy: a platform for interactive large-scale analysis. Genome Research (2005) 15:1451-1455

Taylor, J., et al. Using Galaxy to perform large-scale interactive data analyses. Current Protocols in Bioinformatics (2007) Chapter 10, unit 10.

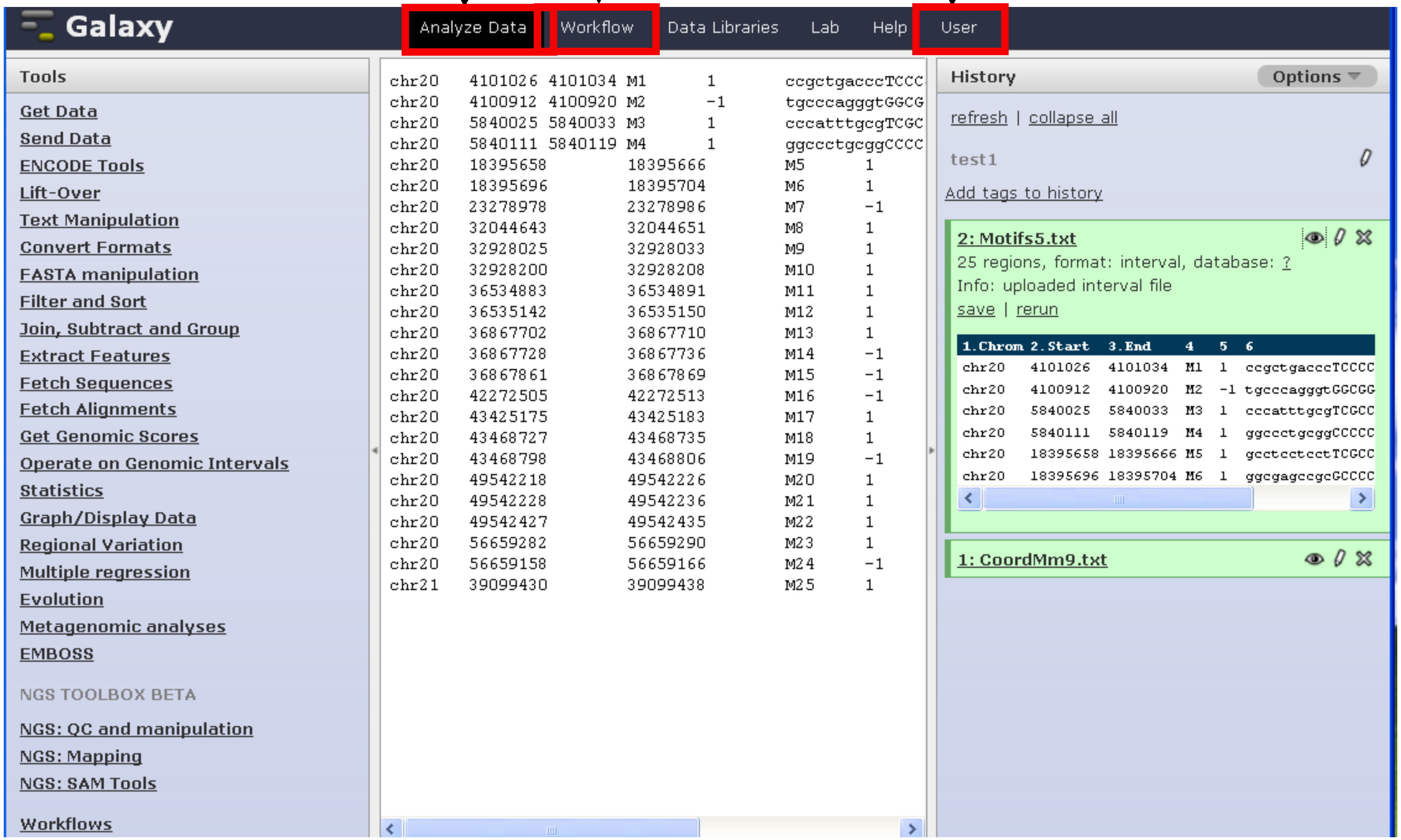
Blankenberg D., et al. Manipulation of FASTQ data with Galaxy. Bioinformatics. 2010 Jul 15;26(14):1783-5

Galaxy Interface

Data analysis

Create analysis pipelines

Log in/out



```
chr20   4101026 4101034 M1      1       ccgctgacccTCCC
chr20   4100912 4100920 M2      -1      tgcccagggtGGCG
chr20   5840025 5840033 M3      1       cccatttgcgTCGC
chr20   5840111 5840119 M4      1       ggccctgcggCCCC
chr20   18395658        18395666        M5      1
chr20   18395696        18395704        M6      1
chr20   23278978        23278986        M7      -1
chr20   32044643        32044651        M8      1
chr20   32928025        32928033        M9      1
chr20   32928200        32928208        M10     1
chr20   36534883        36534891        M11     1
chr20   36535142        36535150        M12     1
chr20   36867702        36867710        M13     1
chr20   36867728        36867736        M14     -1
chr20   36867861        36867869        M15     -1
chr20   42272505        42272513        M16     -1
chr20   43425175        43425183        M17     1
chr20   43468727        43468735        M18     1
chr20   43468798        43468806        M19     -1
chr20   49542218        49542226        M20     1
chr20   49542228        49542236        M21     1
chr20   49542427        49542435        M22     1
chr20   56659282        56659290        M23     1
chr20   56659158        56659166        M24     -1
chr21   39099430        39099438        M25     1
```

2: Motifs5.txt

25 regions, format: interval, database: ?

Info: uploaded interval file

save | rerun

1.Chrom	2.Start	3.End	4	5	6
chr20	4101026	4101034	M1	1	ccgctgacccTCCCC
chr20	4100912	4100920	M2	-1	tgcccagggtGGCGG
chr20	5840025	5840033	M3	1	cccatttgcgTCGCC
chr20	5840111	5840119	M4	1	ggccctgcggCCCCC
chr20	18395658	18395666	M5	1	gcctcctcctTCGCC
chr20	18395696	18395704	M6	1	ggcgagccgcGCCCC

1: CoordMm9.txt

Processed data

Green: job is finished

Yellow: job is running

Gray: job is in queue

Red: there is a problem

Tools window

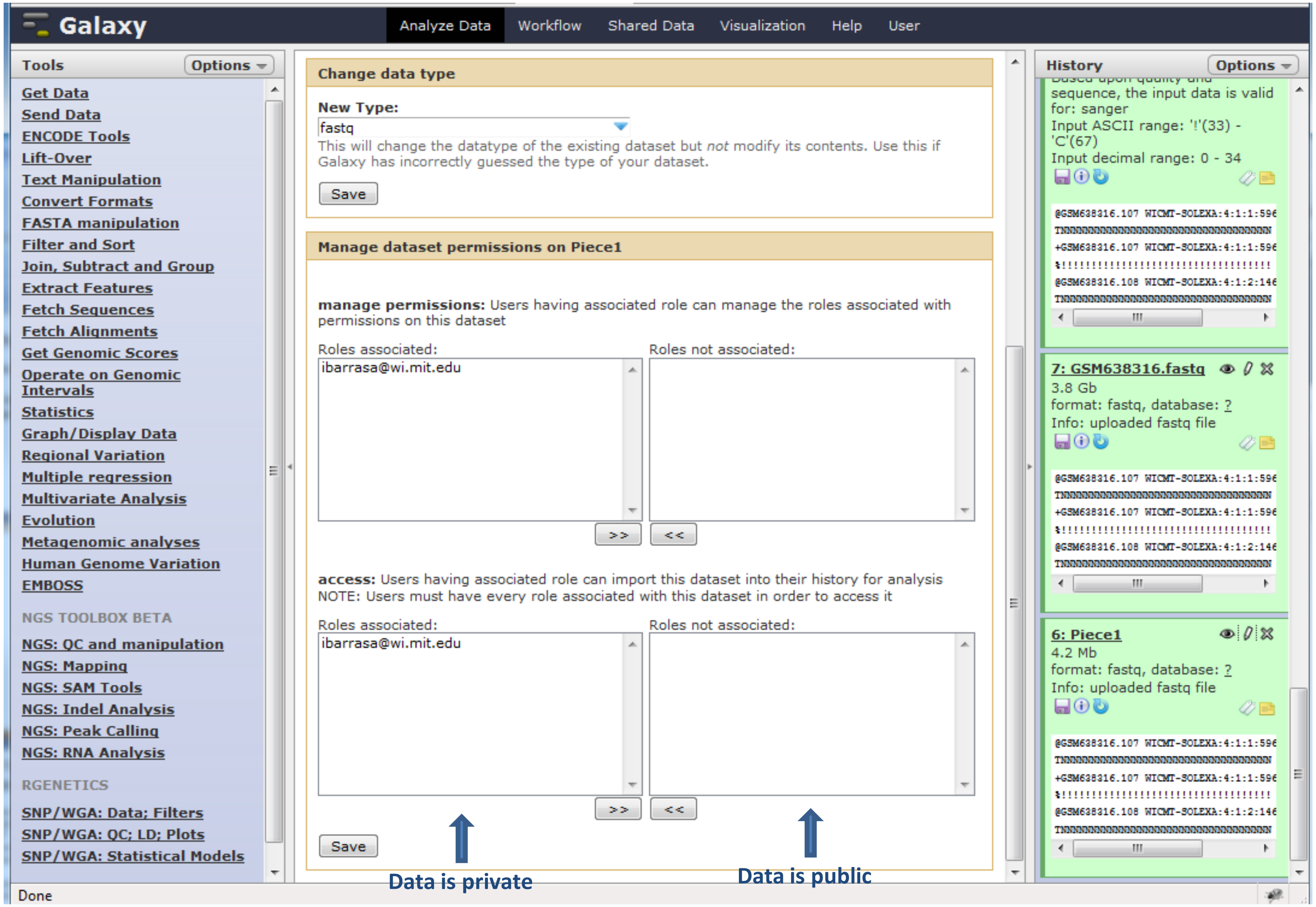
Data display and tools dialog window

History window: datasets for each analysis are kept here

Security issues

- Need to register to be able to keep your data and history (log in button).
- Your data has to be public to be able to be visualized at UCSC. By default the data is public.
- You could make your data private, download it and visualize in UCSC or other browser.

Security issues II



Change data type

New Type: fastq

This will change the datatype of the existing dataset but *not* modify its contents. Use this if Galaxy has incorrectly guessed the type of your dataset.

Manage dataset permissions on Piece1

manage permissions: Users having associated role can manage the roles associated with permissions on this dataset

Roles associated: ibarrasa@wi.mit.edu

Roles not associated:

access: Users having associated role can import this dataset into their history for analysis
NOTE: Users must have every role associated with this dataset in order to access it

Roles associated: ibarrasa@wi.mit.edu

Roles not associated:

Based upon quality and sequence, the input data is valid for: sanger
Input ASCII range: '!'(33) - 'C'(67)
Input decimal range: 0 - 34

```
@GSM638316.107 WICMT-SOLEXA:4:1:1:596
TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+GSM638316.107 WICMT-SOLEXA:4:1:1:596
$!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@GSM638316.108 WICMT-SOLEXA:4:1:2:146
TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

7: GSM638316.fastq
3.8 Gb
format: fastq, database: ?
Info: uploaded fastq file

```
@GSM638316.107 WICMT-SOLEXA:4:1:1:596
TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+GSM638316.107 WICMT-SOLEXA:4:1:1:596
$!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@GSM638316.108 WICMT-SOLEXA:4:1:2:146
TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

6: Piece1
4.2 Mb
format: fastq, database: ?
Info: uploaded fastq file

```
@GSM638316.107 WICMT-SOLEXA:4:1:1:596
TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
+GSM638316.107 WICMT-SOLEXA:4:1:1:596
$!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
@GSM638316.108 WICMT-SOLEXA:4:1:2:146
TNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

Data is private

Data is public

Talk Outline

- Introduction to Galaxy
- Data upload
- Format conversion and quality control tools
- Analysis of ChIP-seq experiments with MACS
- Analysis of RNA-seq experiments with Tuxedo tools
- Demo

Data upload I

- For files larger than 2Gb, transfer to the Galaxy server via the file transfer protocol (FTP).
- Log in to tak (**ssh –l userName tak.wi.mit.edu**), and **cd** to the folder that has your files. (See hot topic “introduction to Unix” http://iona.wi.mit.edu/bio/education/hot_topics/unix_2010/slides.pdf)
- Ftp to Galaxy:

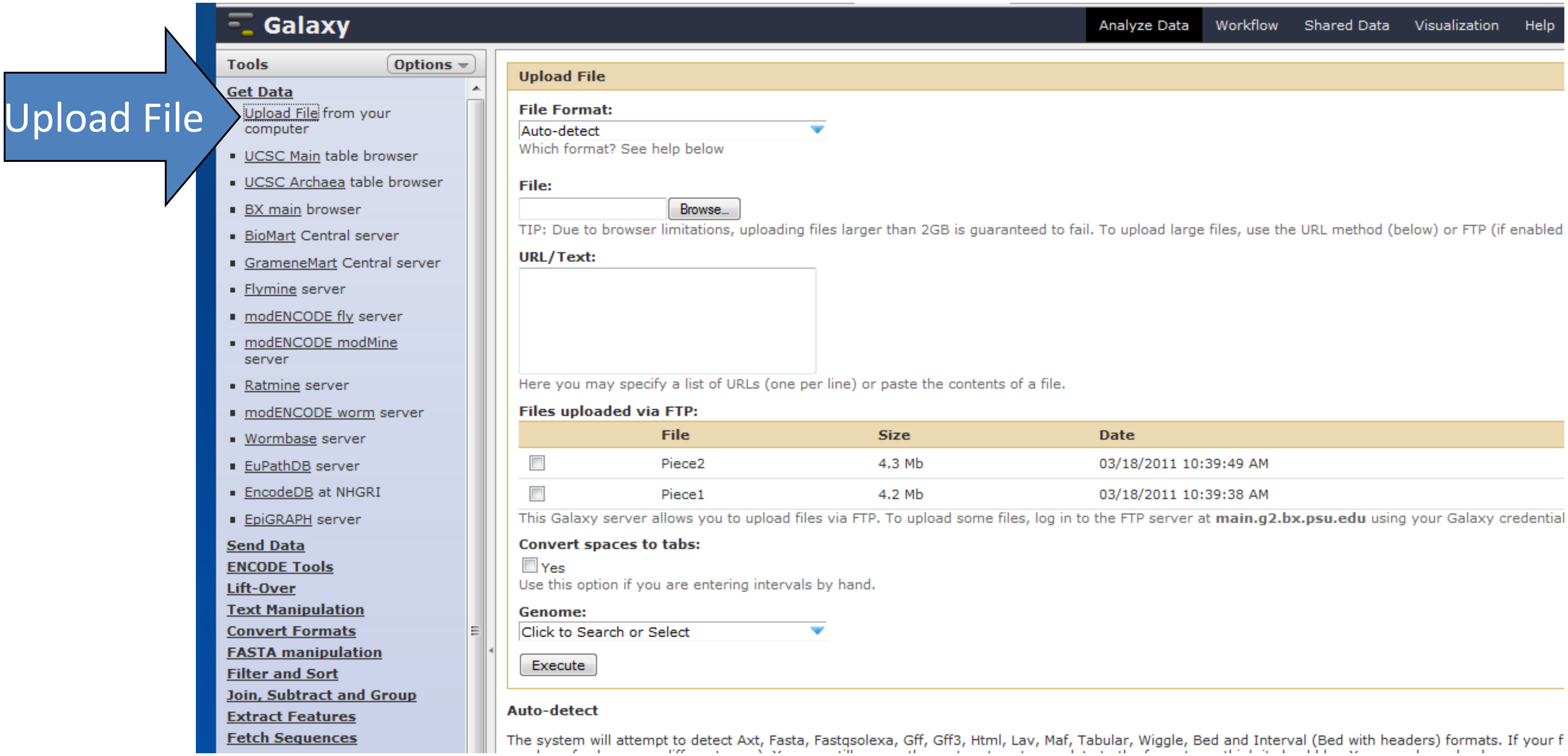
```
ftp main.g2.bx.psu.edu
Name (main.g2.bx.psu.edu:ibarrasa):  Type your email
Password: Type your Galaxy password

230 User ibarrasa@wi.mit.edu logged in
Remote system type is UNIX.
Using binary mode to transfer files.
ftp>
```

- Upload file

```
ftp>  put FileName
ftp> exit
```

Data upload II



Upload File

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser
- BX main browser
- BioMart Central server
- GrameneMart Central server
- Flymine server
- modENCODE fly server
- modENCODE modMine server
- Ratmine server
- modENCODE worm server
- Wormbase server
- EuPathDB server
- EncodeDB at NHGRI
- EpiGRAPH server

Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences

Upload File

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

File:
Browse...
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

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
□	Piece2	4.3 Mb	03/18/2011 10:39:49 AM
□	Piece1	4.2 Mb	03/18/2011 10:39:38 AM

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the FTP server at **main.g2.bx.psu.edu** using your Galaxy credential

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

Genome:
Click to Search or Select

Execute

Auto-detect

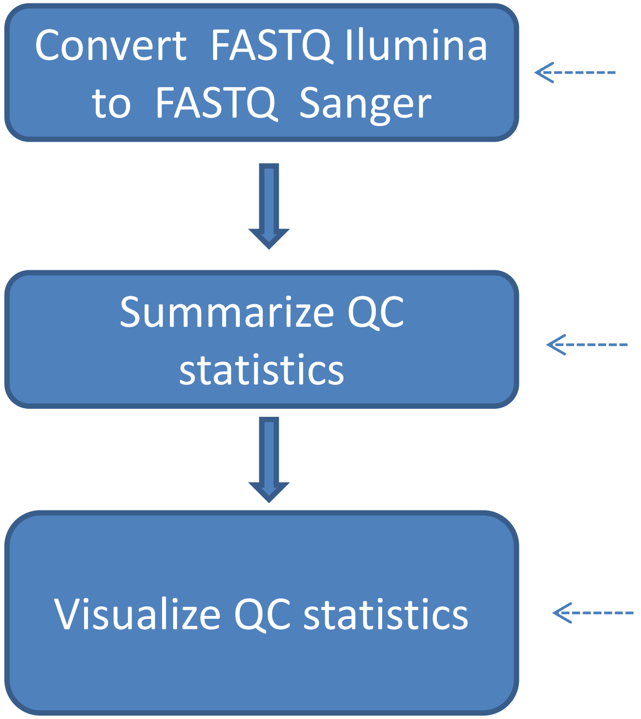
The system will attempt to detect Axt, Fasta, Fastqsolexa, Gff, Gff3, Html, Lav, Maf, Tabular, Wiggle, Bed and Interval (Bed with headers) formats. If your f

Talk Outline

- Introduction to Galaxy
- Data upload
- Format conversion and quality control tools
- Analysis of ChIP-seq experiments with MACS
- Analysis of RNA-seq experiments with Tuxedo tools
- Demo

Format conversion and quality control tools

NGS: QC and manipulation

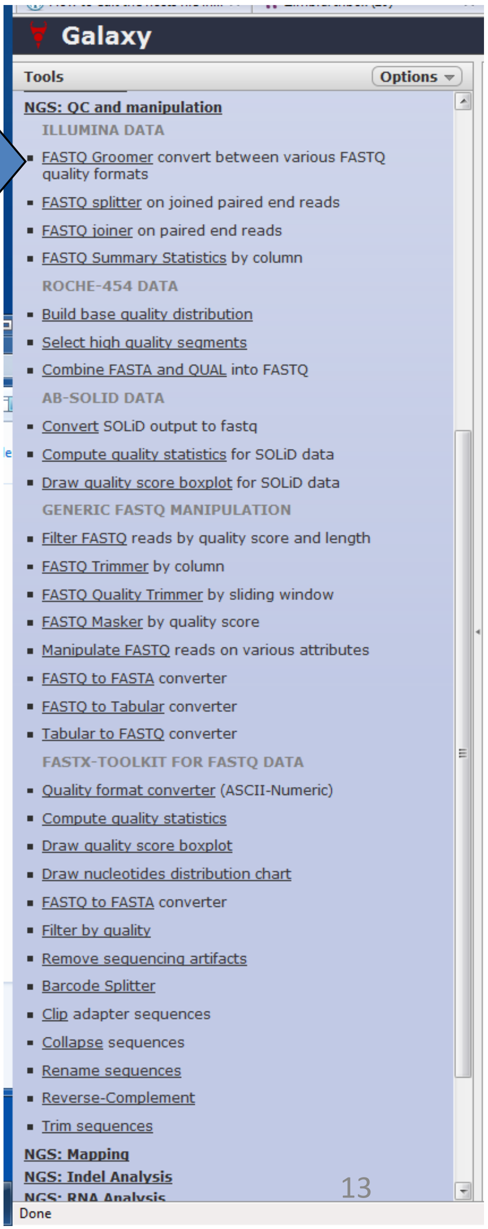


- FASTQ Groomer (convert between various FASTQ quality formats)

- Compute quality statistics

- Draw quality score boxplot
- Draw nucleotides distribution chart

FASTQ Groomer



NGS: QC and manipulation

ILLUMINA DATA

- 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

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score
- Manipulate FASTQ reads on various attributes
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- Tabular to FASTQ converter

FASTX-TOOLKIT FOR FASTQ DATA

- Quality format converter (ASCII-Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter
- Filter by quality
- Remove sequencing artifacts
- Barcode Splitter
- Clip adapter sequences
- Collapse sequences
- Rename sequences
- Reverse-Complement
- Trim sequences

NGS: Mapping

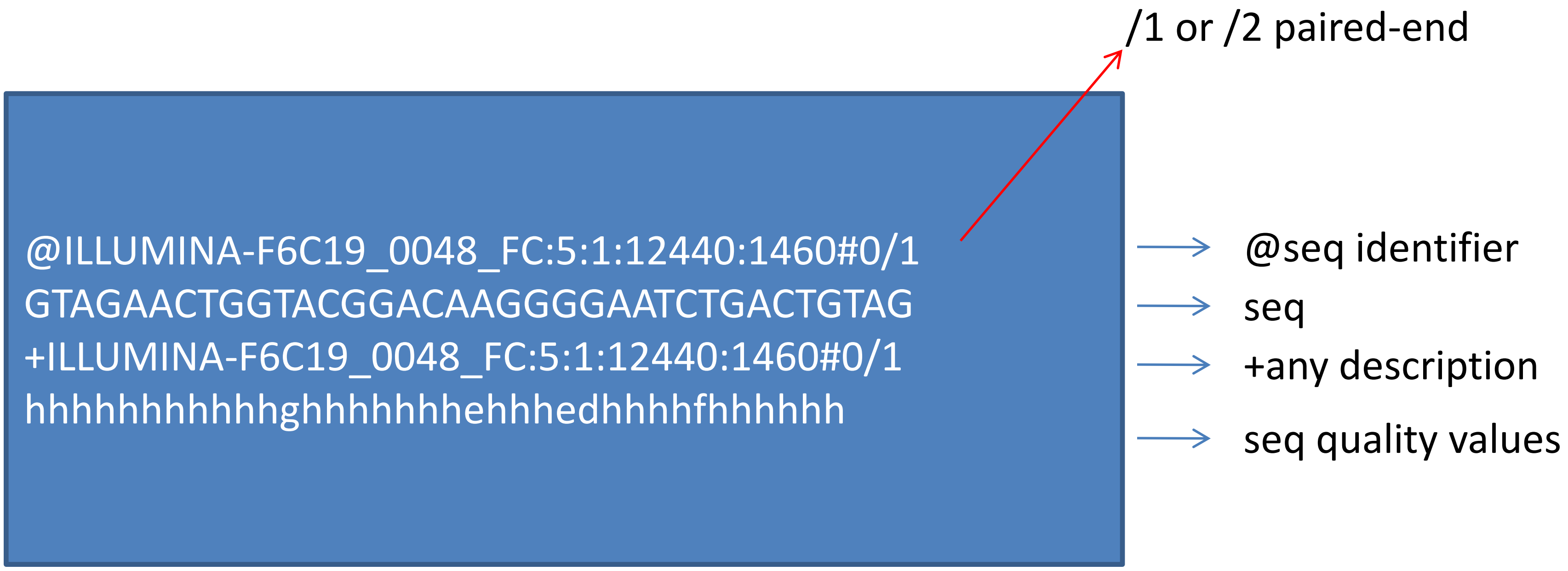
NGS: Indel Analysis

NGS: RNA Analysis

Note: FastQC is not incorporated in Galaxy but it is installed in tak .

Illumina data format

- Fastq format:



/1 or /2 paired-end

```
@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhghhhhhhehhhedhhhhfhhhhhh
```

@seq identifier

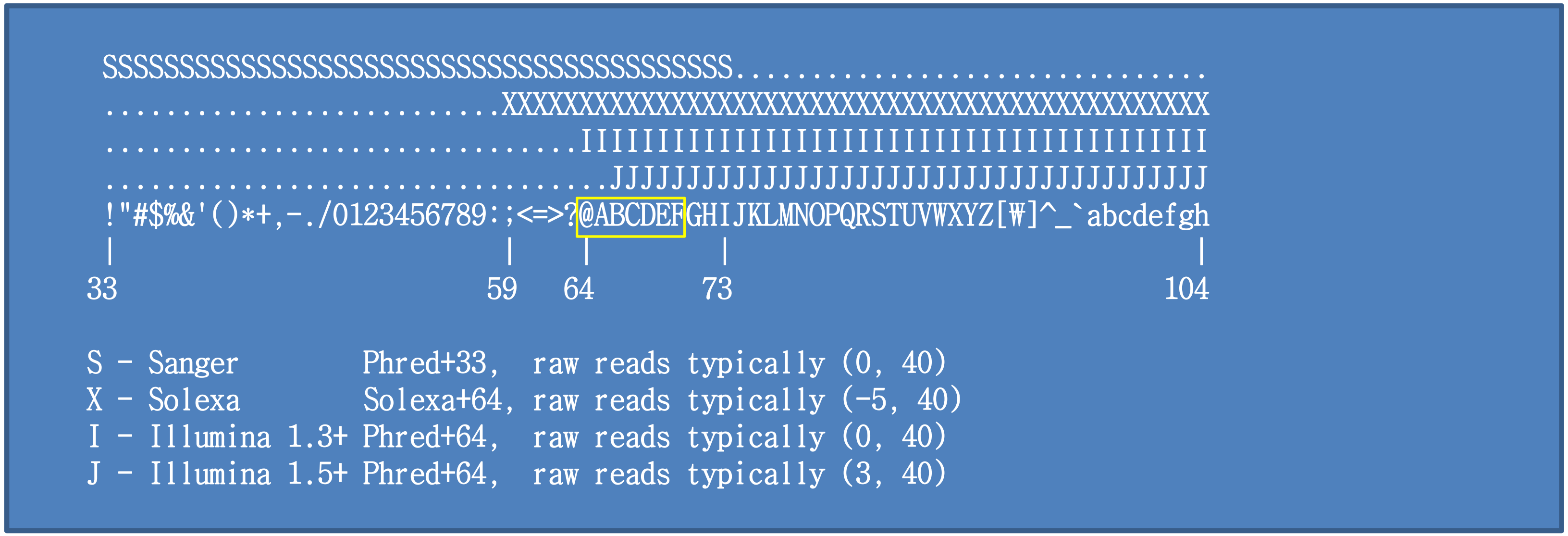
seq

+any description

seq quality values

Sequence quality values on different FASTQ formats

http://en.wikipedia.org/wiki/FASTQ_format

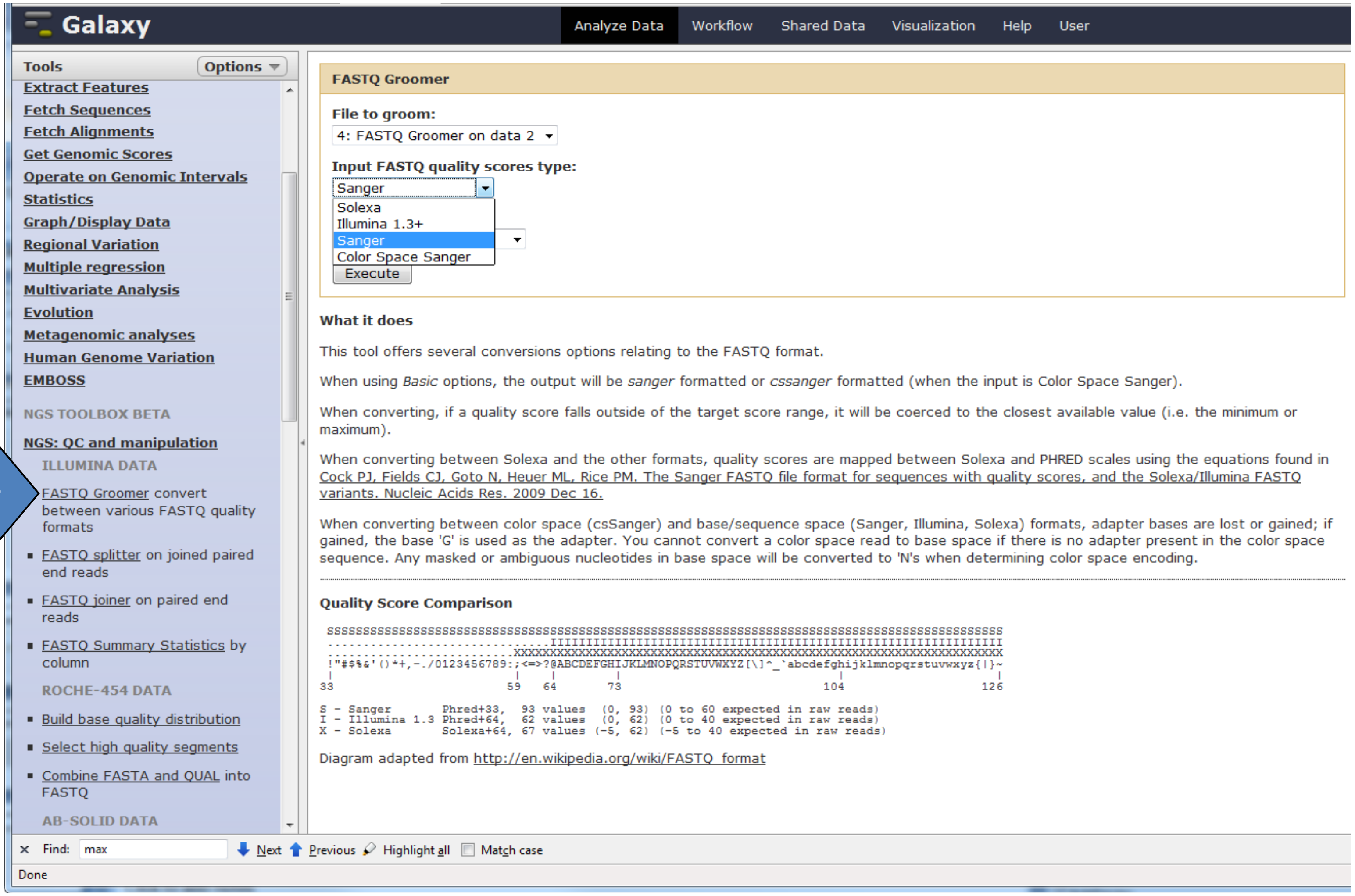


```
  SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS...............................
  ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
  ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
  .................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ
  !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[₩]^_`abcdefgh
  |                         |    |        |                              |
 33                        59   64       73                            104

 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)
```

To discriminate between Solexa and Illumina 1.3+ check if your sequences have any of the characters: :;<=>?

FASTQ Groomer



Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation

ILLUMINA DATA

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

AB-SOLID DATA

FASTQ Groomer

FASTQ Groomer

File to groom: 4: FASTQ Groomer on data 2

Input FASTQ quality scores type: Sanger

Solexa
Illumina 1.3+
Sanger
Color Space Sanger

Execute

What it does

This tool offers several conversions options relating to the FASTQ format.

When using *Basic* options, the output will be *sanger* formatted or *cssanger* formatted (when the input is Color Space Sanger).

When converting, if a quality score falls outside of the target score range, it will be coerced to the closest available value (i.e. the minimum or maximum).

When converting between Solexa and the other formats, quality scores are mapped between Solexa and PHRED scales using the equations found in Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2009 Dec 16.

When converting between color space (csSanger) and base/sequence space (Sanger, Illumina, Solexa) formats, adapter bases are lost or gained; if gained, the base 'G' is used as the adapter. You cannot convert a color space read to base space if there is no adapter present in the color space sequence. Any masked or ambiguous nucleotides in base space will be converted to 'N's when determining color space encoding.

Quality Score Comparison

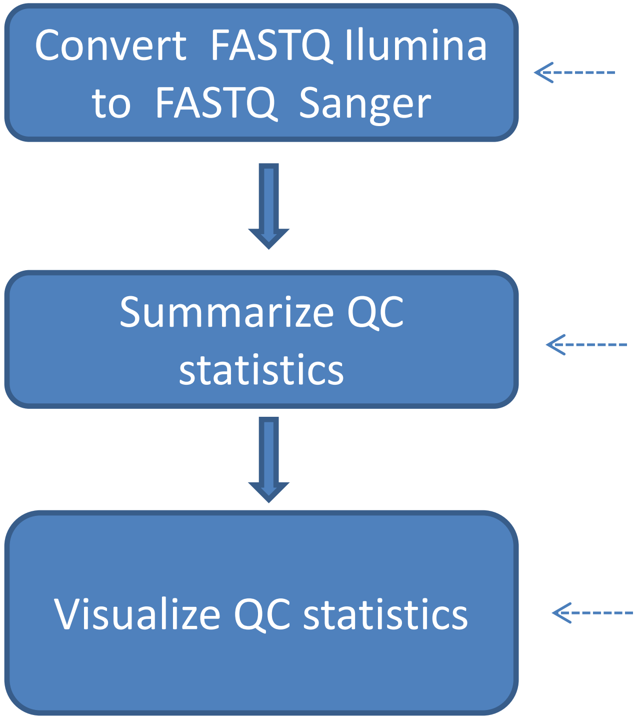
```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
 !"#$%&'()*+,-./0123456789:;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

S - Sanger        Phred+33,  93 values  (0, 93) (0 to 60 expected in raw reads)
I - Illumina 1.3  Phred+64,  62 values  (0, 62) (0 to 40 expected in raw reads)
X - Solexa        Solexa+64, 67 values (-5, 62) (-5 to 40 expected in raw reads)
```

Diagram adapted from http://en.wikipedia.org/wiki/FASTQ_format

Format conversion and quality control tools

NGS: QC and manipulation


Convert FASTQ Ilumina to FASTQ Sanger

Summarize QC statistics

Visualize QC statistics

- FASTQ Groomer (convert between various FASTQ quality formats)

- Compute quality statistics

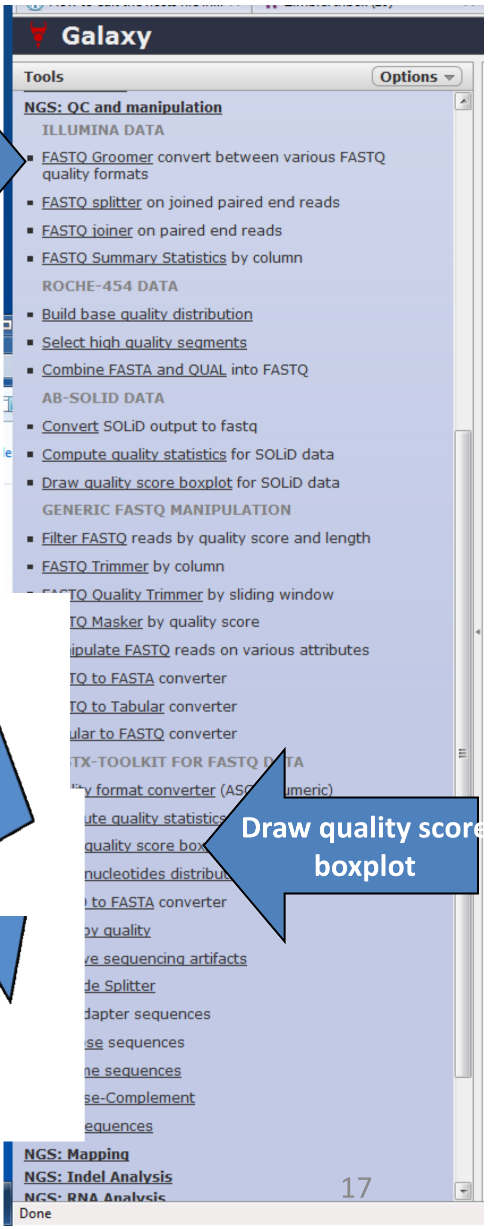
- Draw quality score boxplot
- Draw nucleotides distribution chart

FASTQ Groomer

Compute quality statistics

Draw quality score boxplot

Draw nucleotide distribution chart


Galaxy

Tools Options

NGS: QC and manipulation

ILLUMINA DATA

- 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

AB-SOLID DATA

- Convert SOLiD output to fastq
- Compute quality statistics for SOLiD data
- Draw quality score boxplot for SOLiD data

GENERIC FASTQ MANIPULATION

- Filter FASTQ reads by quality score and length
- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- TQ Masker by quality score
- ipulate FASTQ reads on various attributes
- TQ to FASTA converter
- TQ to Tabular converter
- ular to FASTQ converter

TX-TOOLKIT FOR FASTQ D TA

- ity format converter (AS meric)
- ute quality statistics
- quality score box
- nucleotides distribu
- to FASTA converter
- by quality
- ve sequencing artifacts
- de Splitter
- dapter sequences
- se sequences
- ne sequences
- se-Complement
- equences

NGS: Mapping

NGS: Indel Analysis

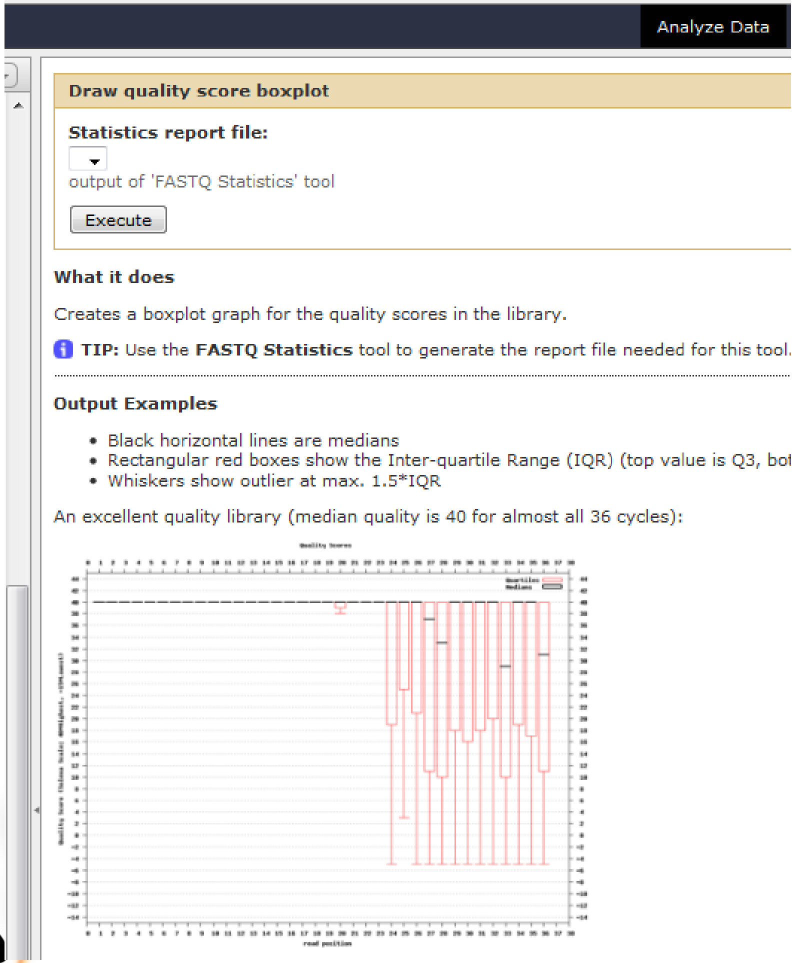
NGS: RNA Analysis

Done

Note: FastQC is not incorporated in Galaxy but it is installed in tak .

Quality control visualization tools

Draw quality score boxplot



Draw quality score boxplot

Statistics report file:

output of 'FASTQ Statistics' tool

What it does

Creates a boxplot graph for the quality scores in the library.

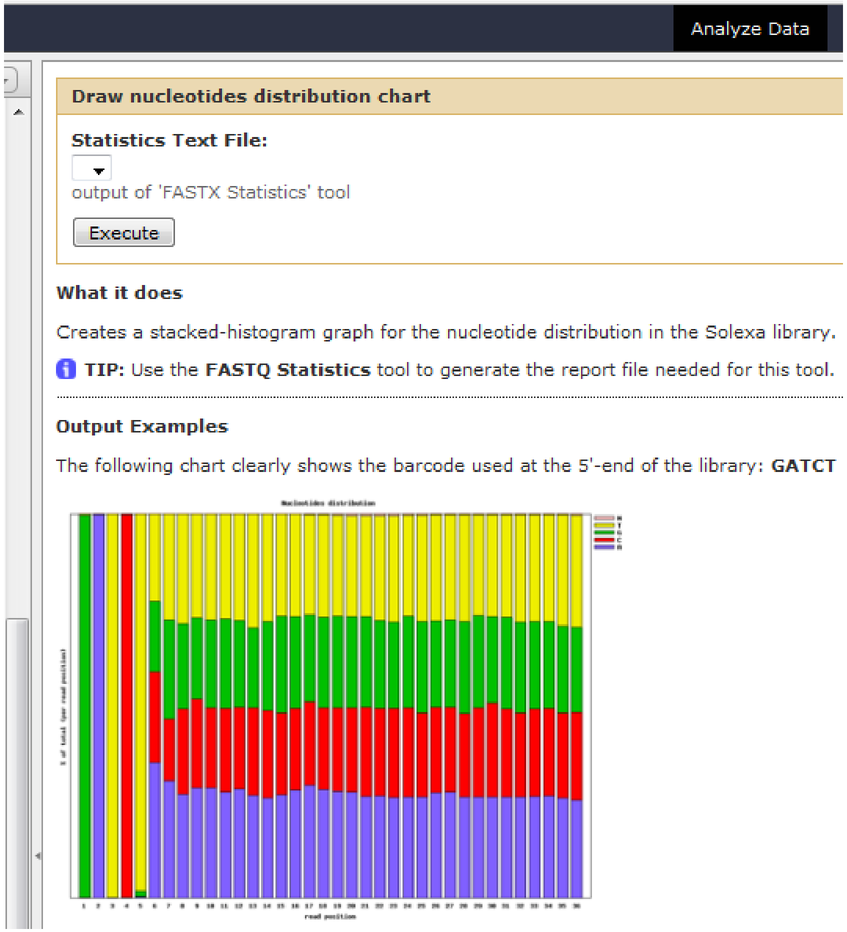
TIP: Use the **FASTQ Statistics** tool to generate the report file needed for this tool.

Output Examples

- Black horizontal lines are medians
- Rectangular red boxes show the Inter-quartile Range (IQR) (top value is Q3, bot
- Whiskers show outlier at max. 1.5*IQR

An excellent quality library (median quality is 40 for almost all 36 cycles):

Draw nucleotides distribution chart



Draw nucleotides distribution chart

Statistics Text File:

output of 'FASTX Statistics' tool

What it does

Creates a stacked-histogram graph for the nucleotide distribution in the Solexa library.

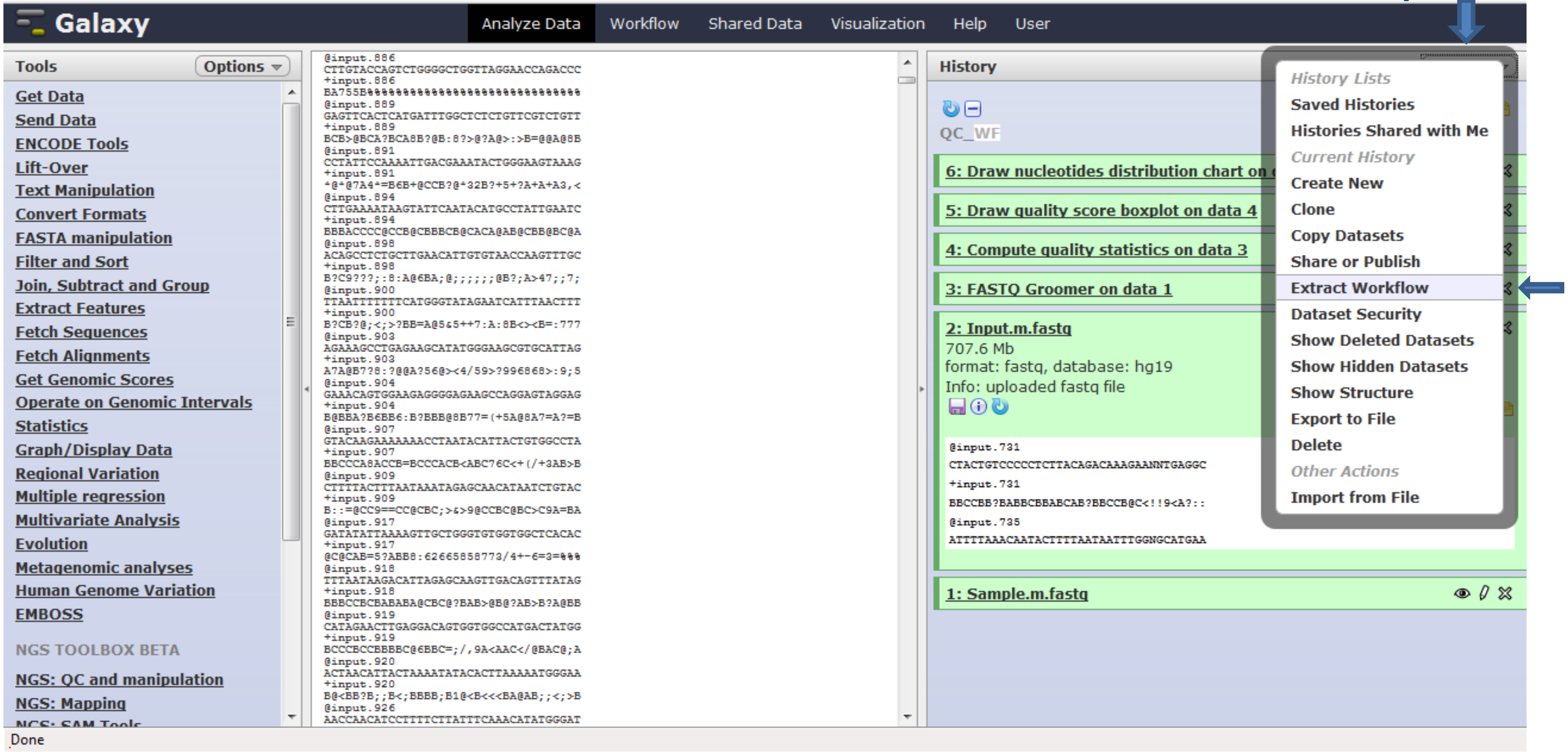
TIP: Use the **FASTQ Statistics** tool to generate the report file needed for this tool.

Output Examples

The following chart clearly shows the barcode used at the 5'-end of the library: **GATCT**

How to make a workflow from the history

History Options

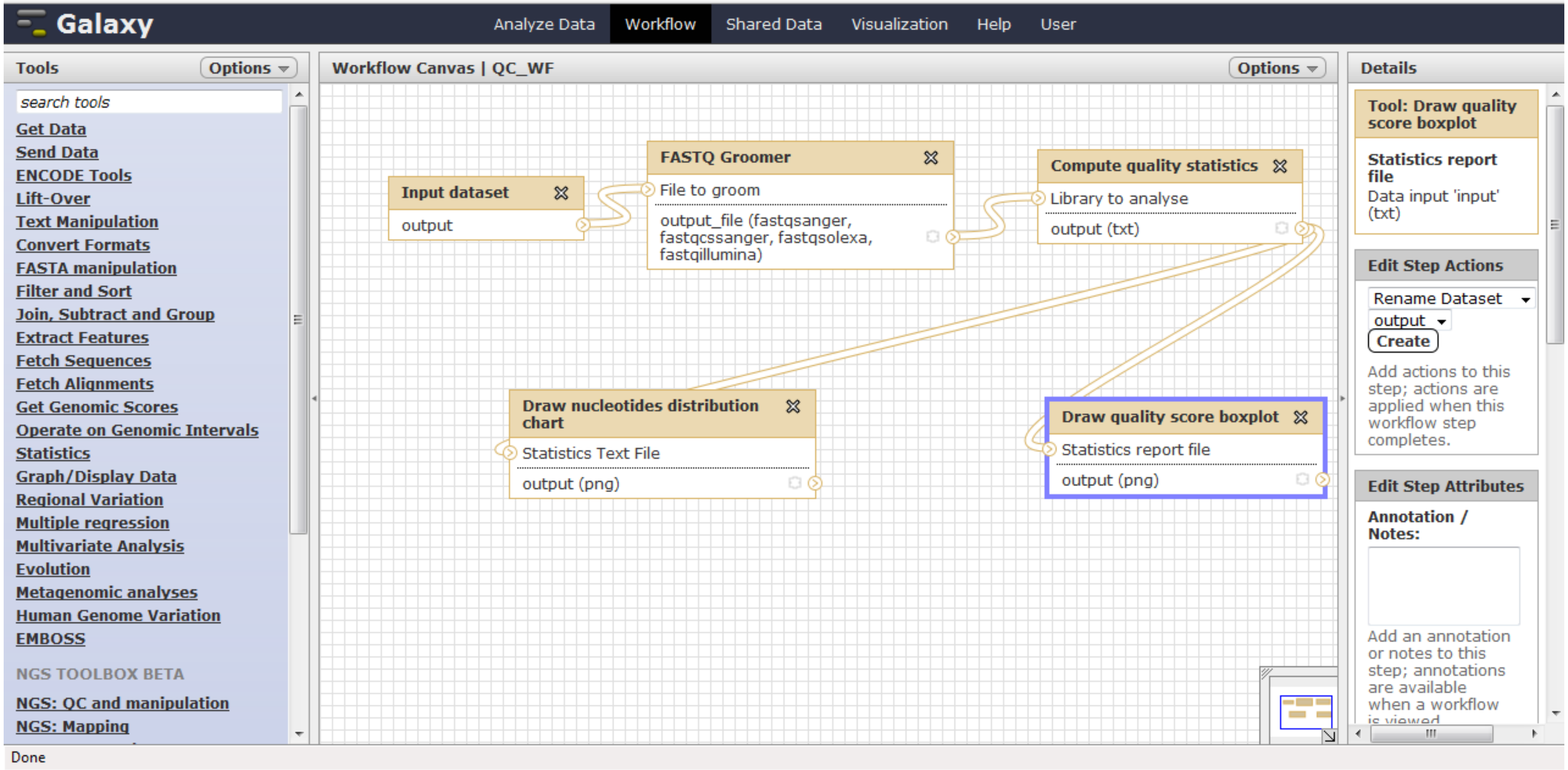


```
@input.886
CTTGTACCAGTCTGGGGCTGGTTAGGAACCAGACCC
+input.886
BA755B%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@input.889
GAGTTCACTCATGATTTGGCTCTCTGTTCGTCTGTT
+input.889
BCB>@BCA?BCA8B?@B:8?>@?A@>:>B=@@A@8B
@input.891
CCTATTCCAAAATTGACGAAATACTGGGAAGTAAAG
+input.891
*@*@7A4*=B6B+@CCB?@*32B?+5+?A+A+A3,<
@input.894
CTTGAAAATAAGTATTCAATACATGCCTATTGAATC
+input.894
BBBACCCC@CCB@CBBBCB@CACA@AB@CBB@BC@A
@input.898
ACAGCCTCTGCTTGAACATTGTGTAACCAAGTTTGC
+input.898
B?C9???;:8:A@6BA;@;;;;;;@B?;A>47;;7;
@input.900
TTAATTTTTTTCATGGGTATAGAATCATTTAACTTT
+input.900
B?CB?@;<;>?BB=A@5&5++7:A:8B<><B=:777
@input.903
AGAAAGCCTGAGAAGCATATGGGAAGCGTGCATTAG
+input.903
A7A@B7?8:?@@A?56@><4/59>?996868>:9;5
@input.904
GAAACAGTGGAAGAGGGGAGAAGCCAGGAGTAGGAG
+input.904
B@BBA?B6BB6:B?BBB@8B77=(+5A@8A7=A?=B
@input.907
GTACAAGAAAAAAACCTAATACATTACTGTGGCCTA
+input.907
BBCCCA8ACCB=BCCCACB<ABC76C<+(/+3AB>B
@input.909
CTTTTACTTTAATAAATAGAGCAACATAATCTGTAC
+input.909
B::=@CC9==CC@CBC;>&>9@CCBC@BC>C9A=BA
@input.917
GATATATTAAAAGTTGCTGGGTGTGGTGGCTCACAC
+input.917
@C@CAB=5?ABB8:62665858773/4+-6=3=%%%
@input.918
TTTAATAAGACATTAGAGCAAGTTGACAGTTTATAG
+input.918
BBBCCBCBABABA@CBC@?BAB>@B@?AB>B?A@BB
@input.919
CATAGAACTTGAGGACAGTGGTGGCCATGACTATGG
+input.919
BCCCBCCBBBBC@6BBC=;/,9A<AAC</@BAC@;A
@input.920
ACTAACATTACTAAAATATACACTTAAAAATGGGAA
+input.920
B@<BB?B;;B<;BBBB;B1@<B<<<BA@AB;;<;>B
@input.926
AACCAACATCCTTTTCTTATTTCAAACATATGGGAT
```

2: Input.m.fastq
707.6 Mb
format: fastq, database: hg19
Info: uploaded fastq file

```
@input.731
CTACTGTCCCCCTCTTACAGACAAAGAANNTGAGGC
+input.731
BBCCBB?BABBCBBABCAB?BBCCB@C<!!9<A?::
@input.735
ATTTTAAACAATACTTTTAATAATTTGGNGCATGAA
```

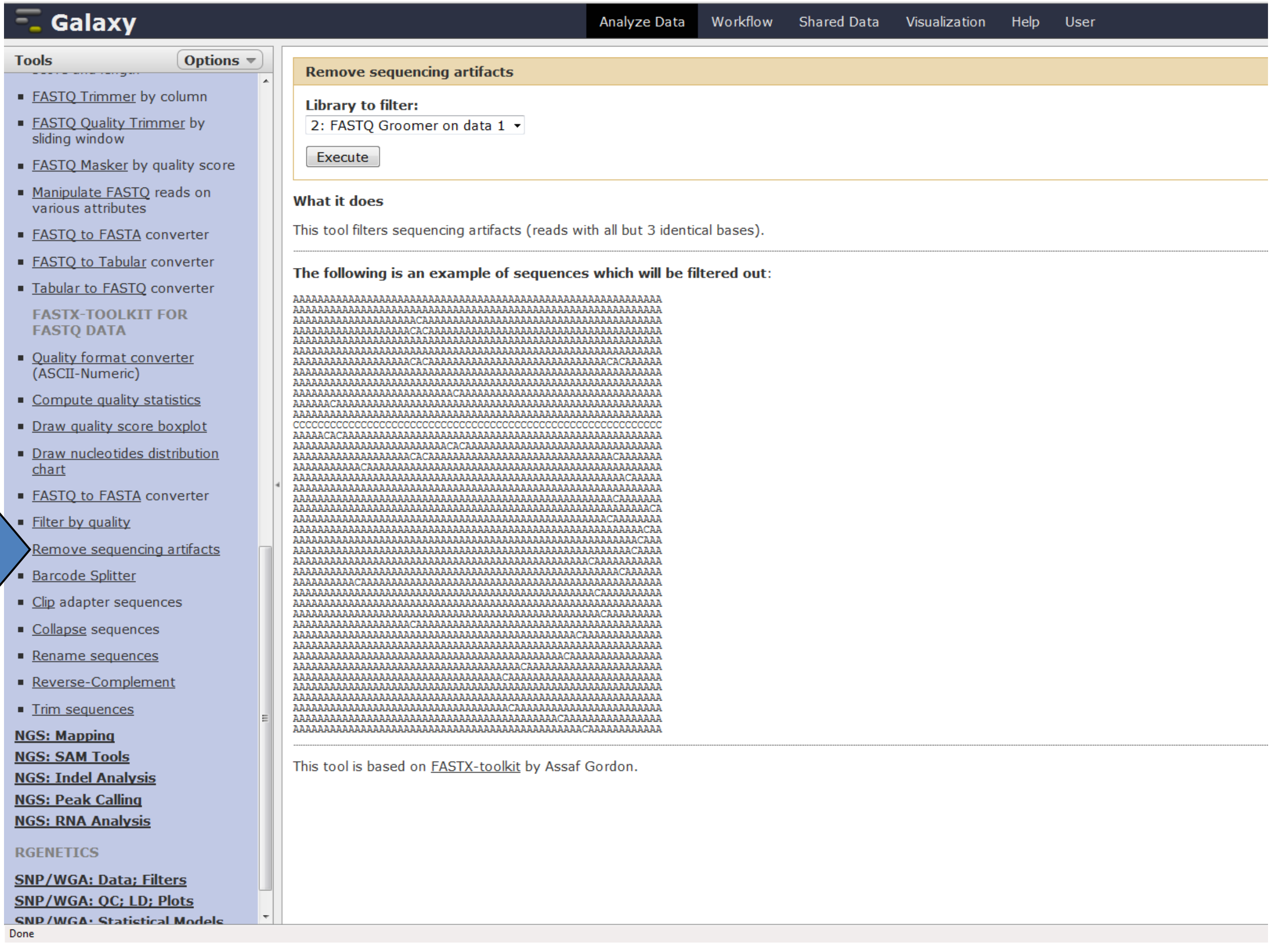
Workflow for Quality Control



Add actions to this step; actions are applied when this workflow step completes.

Add an annotation or notes to this step; annotations are available when a workflow is viewed

Remove sequencing artifacts



- FASTQ Trimmer by column
- FASTQ Quality Trimmer by sliding window
- FASTQ Masker by quality score
- Manipulate FASTQ reads on various attributes
- FASTQ to FASTA converter
- FASTQ to Tabular converter
- Tabular to FASTQ converter

FASTX-TOOLKIT FOR FASTQ DATA

- Quality format converter (ASCII-Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter
- Filter by quality
- Remove sequencing artifacts
- Barcode Splitter
- Clip adapter sequences
- Collapse sequences
- Rename sequences
- Reverse-Complement
- Trim sequences

NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models

Done

Remove sequencing artifacts

Library to filter:
2: FASTQ Groomer on data 1

Execute

What it does

This tool filters sequencing artifacts (reads with all but 3 identical bases).

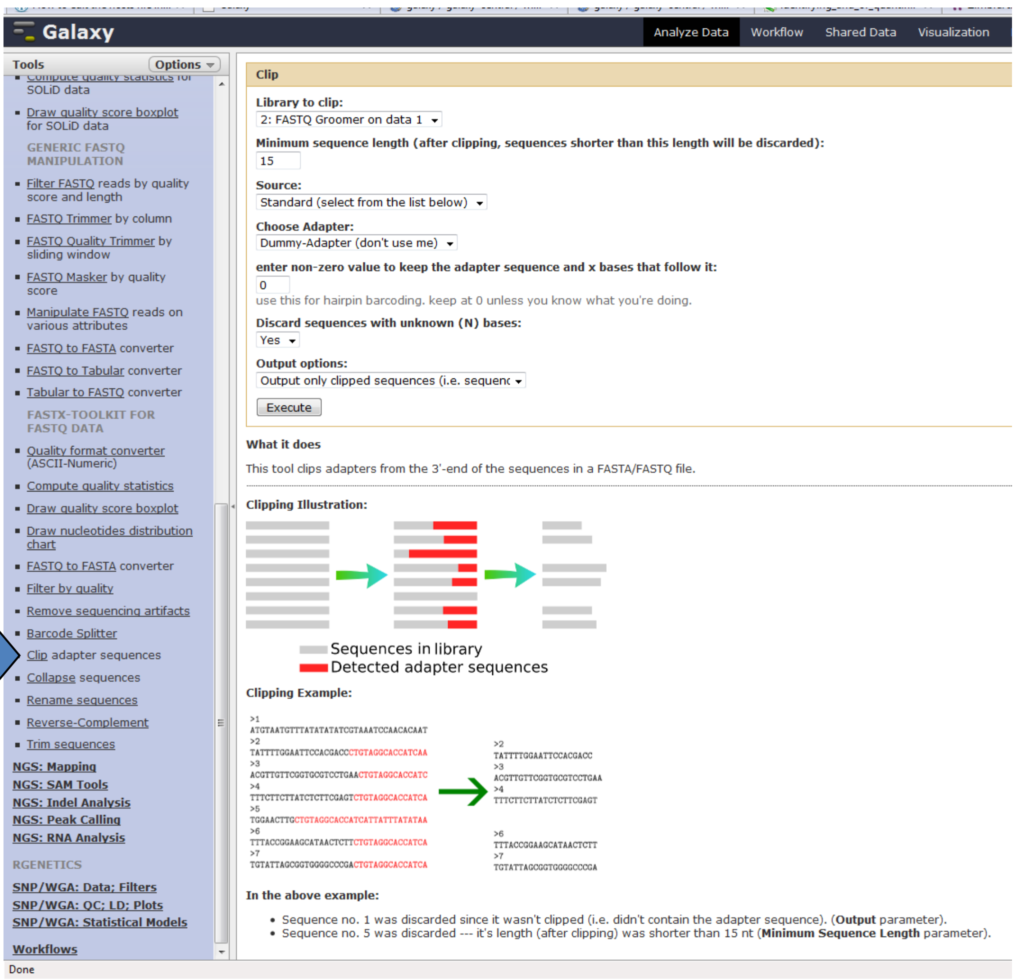
The following is an example of sequences which will be filtered out:

```
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAACACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAACACAAAAAAAAAAAAAAAAAAAAAAAAAAAAACACAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC
AAAAACACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAACACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAACACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAA
AAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAA
AAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAA
AAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAACAAAAAAAAAAAA
```

This tool is based on FASTX-toolkit by Assaf Gordon.

Remove sequencing artifacts

Clip adapter sequences


Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Tools
Options
Compute quality statistics for SOLiD data
Draw quality score boxplot for SOLiD data
GENERIC FASTQ MANIPULATION
Filter FASTQ reads by quality score and length
FASTQ Trimmer by column
FASTQ Quality Trimmer by sliding window
FASTQ Masker by quality score
Manipulate FASTQ reads on various attributes
FASTQ to FASTA converter
FASTQ to Tabular converter
Tabular to FASTQ converter
FASTX-TOOLKIT FOR FASTQ DATA
Quality format converter (ASCII-Numeric)
Compute quality statistics
Draw quality score boxplot
Draw nucleotides distribution chart
FASTQ to FASTA converter
Filter by quality
Remove sequencing artifacts
Barcode Splitter
Clip adapter sequences
Collapse sequences
Rename sequences
Reverse-Complement
Trim sequences
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis
RGENETICS
SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models
Workflows
Done
Execute

Clip

Library to clip:
2: FASTQ Groomer on data 1

Minimum sequence length (after clipping, sequences shorter than this length will be discarded):
15

Source:
Standard (select from the list below)

Choose Adapter:
Dummy-Adapter (don't use me)

enter non-zero value to keep the adapter sequence and x bases that follow it:
0
use this for hairpin barcoding. keep at 0 unless you know what you're doing.

Discard sequences with unknown (N) bases:
Yes

Output options:
Output only clipped sequences (i.e. sequenc

What it does

This tool clips adapters from the 3'-end of the sequences in a FASTA/FASTQ file.

Clipping Illustration:

▭ Sequences in library
🟥 Detected adapter sequences

Clipping Example:

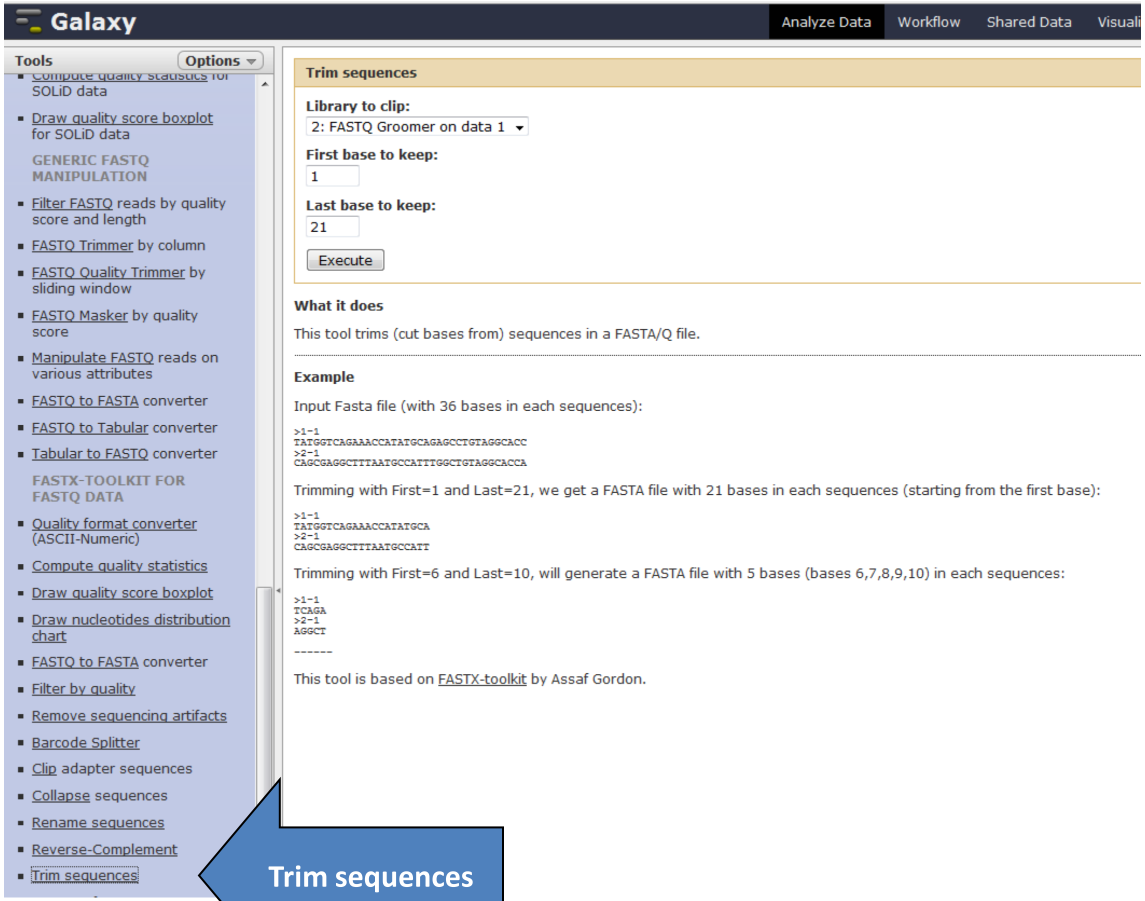
```
>1
ATGTAATGTTTATATATATCGTAAATCCAACACAAT
>2                                         >2
TATTTTGGAATTCCACGACCCTGTAGGCACCATCAA       TATTTTGGAATTCCACGACC
>3                                         >3
ACGTTGTTCGGTGCGTCCTGAACTGTAGGCACCATC       ACGTTGTTCGGTGCGTCCTGAA
>4                                    →    >4
TTTCTTCTTATCTCTTCGAGTCTGTAGGCACCATCA       TTTCTTCTTATCTCTTCGAGT
>5
TGGAACTTGCTGTAGGCACCATCATTATTTATATAA
>6                                         >6
TTTACCGGAAGCATAACTCTTCTGTAGGCACCATCA       TTTACCGGAAGCATAACTCTT
>7                                         >7
TGTATTAGCGGTGGGGCCCGACTGTAGGCACCATCA       TGTATTAGCGGTGGGGCCCGA
```

In the above example:

- Sequence no. 1 was discarded since it wasn't clipped (i.e. didn't contain the adapter sequence). (**Output** parameter).
- Sequence no. 5 was discarded --- it's length (after clipping) was shorter than 15 nt (**Minimum Sequence Length** parameter).

Clip adapter sequences

Trim sequences


Galaxy
Analyze Data
Workflow
Shared Data
Visuali
Tools
Options
Compute quality statistics for SOLiD data
Draw quality score boxplot for SOLiD data
GENERIC FASTQ MANIPULATION
Filter FASTQ reads by quality score and length
FASTQ Trimmer by column
FASTQ Quality Trimmer by sliding window
FASTQ Masker by quality score
Manipulate FASTQ reads on various attributes
FASTQ to FASTA converter
FASTQ to Tabular converter
Tabular to FASTQ converter
FASTX-TOOLKIT FOR FASTQ DATA
Quality format converter (ASCII-Numeric)
Compute quality statistics
Draw quality score boxplot
Draw nucleotides distribution chart
FASTQ to FASTA converter
Filter by quality
Remove sequencing artifacts
Barcode Splitter
Clip adapter sequences
Collapse sequences
Rename sequences
Reverse-Complement
Trim sequences
Execute

Trim sequences

Library to clip: 2: FASTQ Groomer on data 1

First base to keep: 1

Last base to keep: 21

What it does

This tool trims (cut bases from) sequences in a FASTA/Q file.

Example

Input Fasta file (with 36 bases in each sequences):

```
>1-1
TATGGTCAGAAACCATATGCAGAGCCTGTAGGCACC
>2-1
CAGCGAGGCTTTAATGCCATTTGGCTGTAGGCACCA
```

Trimming with First=1 and Last=21, we get a FASTA file with 21 bases in each sequences (starting from the first base):

```
>1-1
TATGGTCAGAAACCATATGCA
>2-1
CAGCGAGGCTTTAATGCCATT
```

Trimming with First=6 and Last=10, will generate a FASTA file with 5 bases (bases 6,7,8,9,10) in each sequences:

```
>1-1
TCAGA
>2-1
AGGCT
```

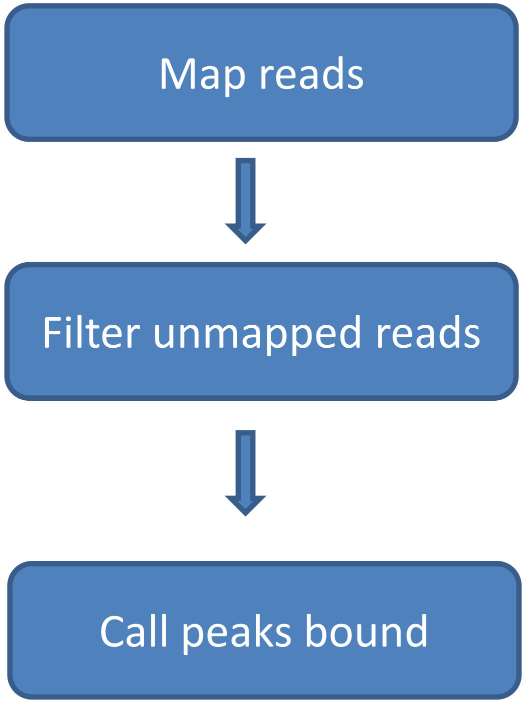
This tool is based on FASTX-toolkit by Assaf Gordon.

Trim sequences

Talk Outline

- Introduction to Galaxy
- Data upload
- Format conversion and quality control tools
- Analysis of ChIP-seq experiments with MACS
- Analysis of RNA-seq experiments with Tuxedo tools
- Demo

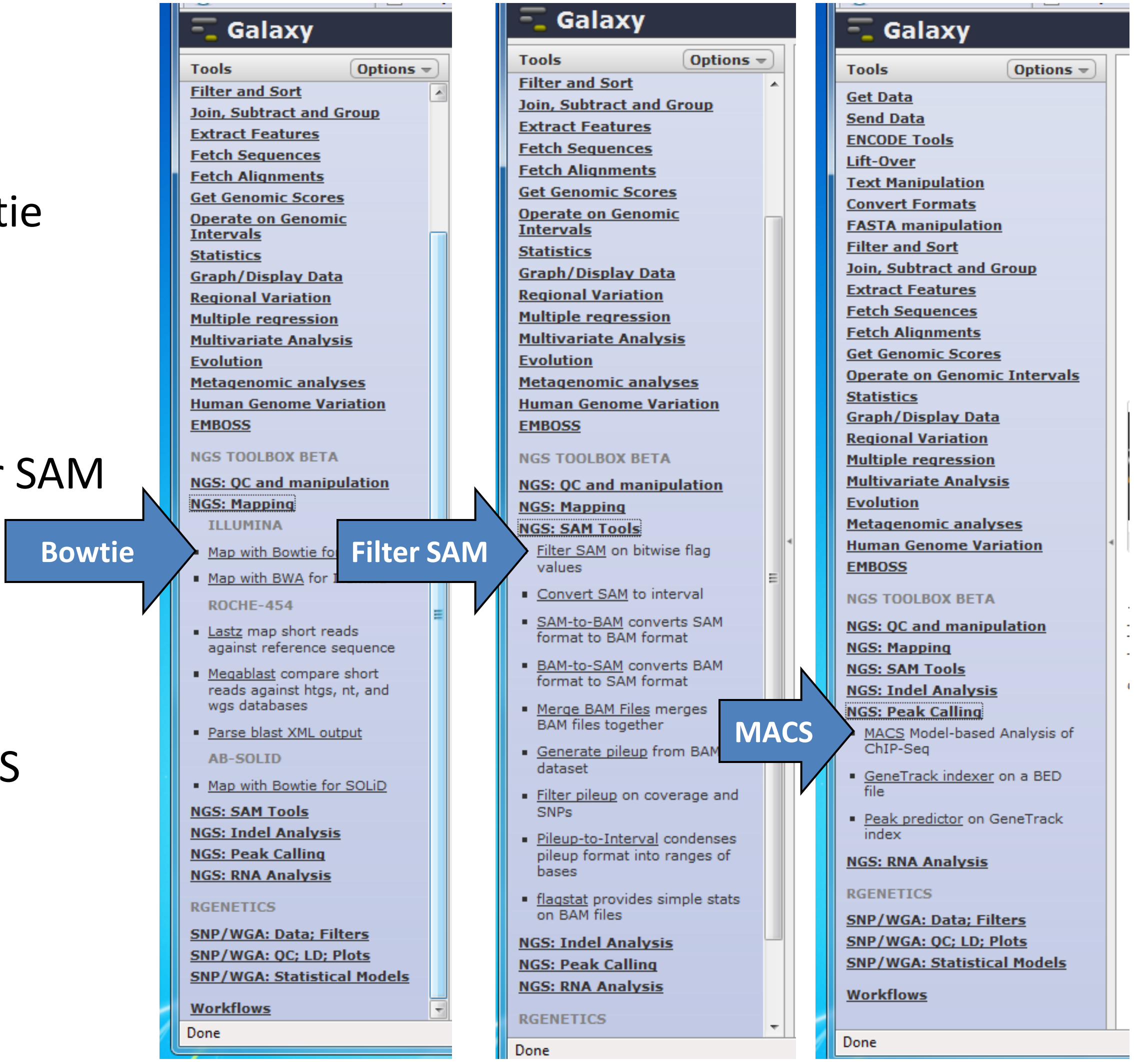
Analysis of ChIP-seq experiments



Bowtie

Filter SAM

MACS



Galaxy

Tools Options

Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping

ILLUMINA

- Map with Bowtie fo
- Map with BWA for I

ROCHE-454

- Lastz map short reads against reference sequence
- Megablast compare short reads against htgs, nt, and wgs databases
- Parse blast XML output

AB-SOLID

- Map with Bowtie for SOLiD

NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis

RGENETICS

SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models

Workflows

Done

Galaxy

Tools Options

Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools

- Filter SAM on bitwise flag values
- Convert SAM to interval
- SAM-to-BAM converts SAM format to BAM format
- BAM-to-SAM converts BAM format to SAM format
- Merge BAM Files merges BAM files together
- Generate pileup from BAM dataset
- Filter pileup on coverage and SNPs
- Pileup-to-Interval condenses pileup format into ranges of bases
- flagstat provides simple stats on BAM files

NGS: Indel Analysis
NGS: Peak Calling
NGS: RNA Analysis

RGENETICS

Done

Galaxy

Tools Options

Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Convert Formats
FASTA manipulation
Filter and Sort
Join, Subtract and Group
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Human Genome Variation
EMBOSS

NGS TOOLBOX BETA

NGS: QC and manipulation
NGS: Mapping
NGS: SAM Tools
NGS: Indel Analysis
NGS: Peak Calling

- MACS Model-based Analysis of ChIP-Seq
- GeneTrack indexer on a BED file
- Peak predictor on GeneTrack index

NGS: RNA Analysis

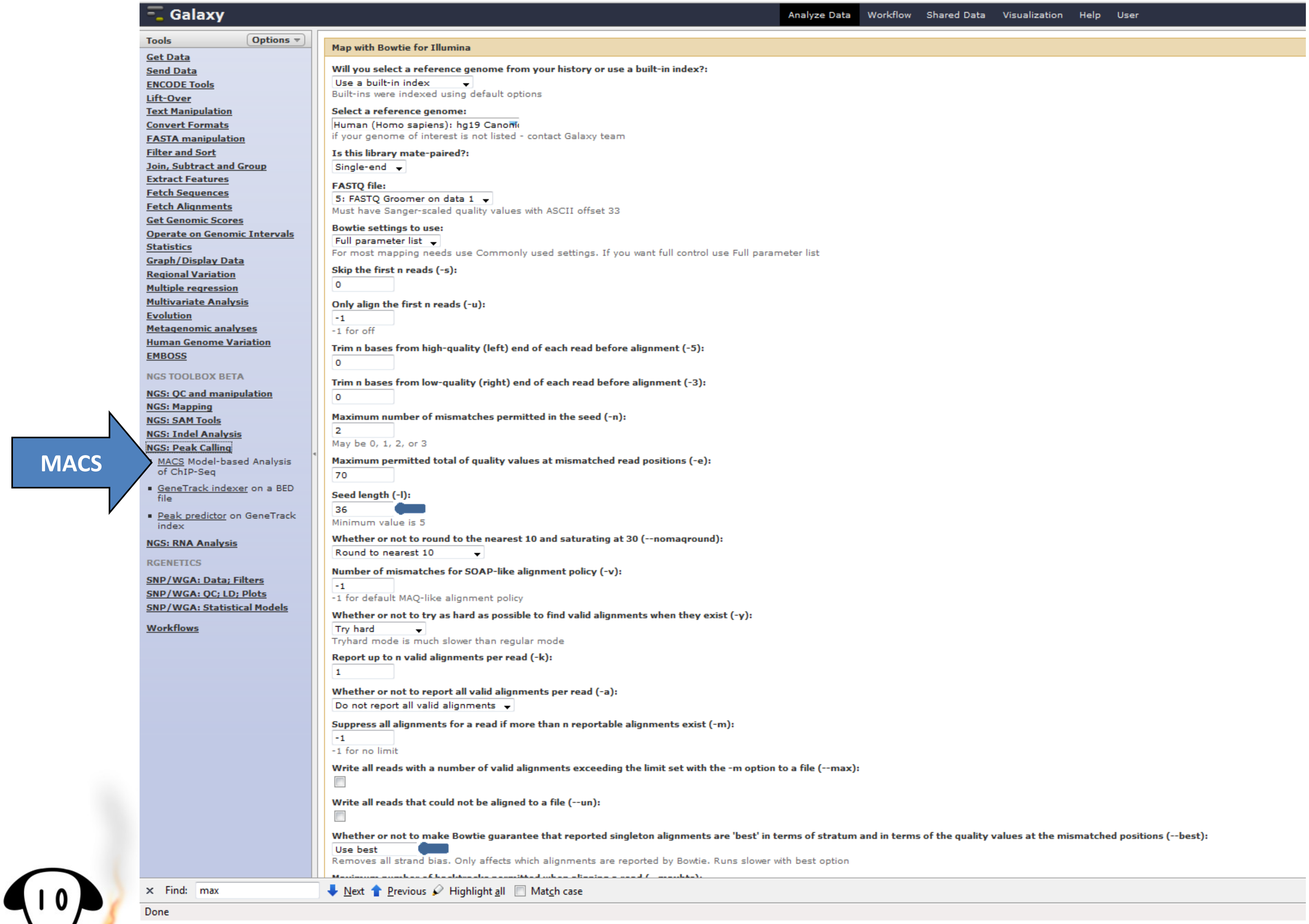
RGENETICS

SNP/WGA: Data; Filters
SNP/WGA: QC; LD; Plots
SNP/WGA: Statistical Models

Workflows

Done

Mapping Reads with Bowtie



MACS

Map with Bowtie for Illumina

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:
Human (Homo sapiens): hg19 Canoni
if your genome of interest is not listed - contact Galaxy team

Is this library mate-paired?:
Single-end

FASTQ file:
5: FASTQ Groomer on data 1
Must have Sanger-scaled quality values with ASCII offset 33

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

Skip the first n reads (-s):
0

Only align the first n reads (-u):
-1
-1 for off

Trim n bases from high-quality (left) end of each read before alignment (-5):
0

Trim n bases from low-quality (right) end of each read before alignment (-3):
0

Maximum number of mismatches permitted in the seed (-n):
2
May be 0, 1, 2, or 3

Maximum permitted total of quality values at mismatched read positions (-e):
70

Seed length (-l):
36
Minimum value is 5

Whether or not to round to the nearest 10 and saturating at 30 (--nomaqround):
Round to nearest 10

Number of mismatches for SOAP-like alignment policy (-v):
-1
-1 for default MAQ-like alignment policy

Whether or not to try as hard as possible to find valid alignments when they exist (-y):
Try hard
Tryhard mode is much slower than regular mode

Report up to n valid alignments per read (-k):
1

Whether or not to report all valid alignments per read (-a):
Do not report all valid alignments

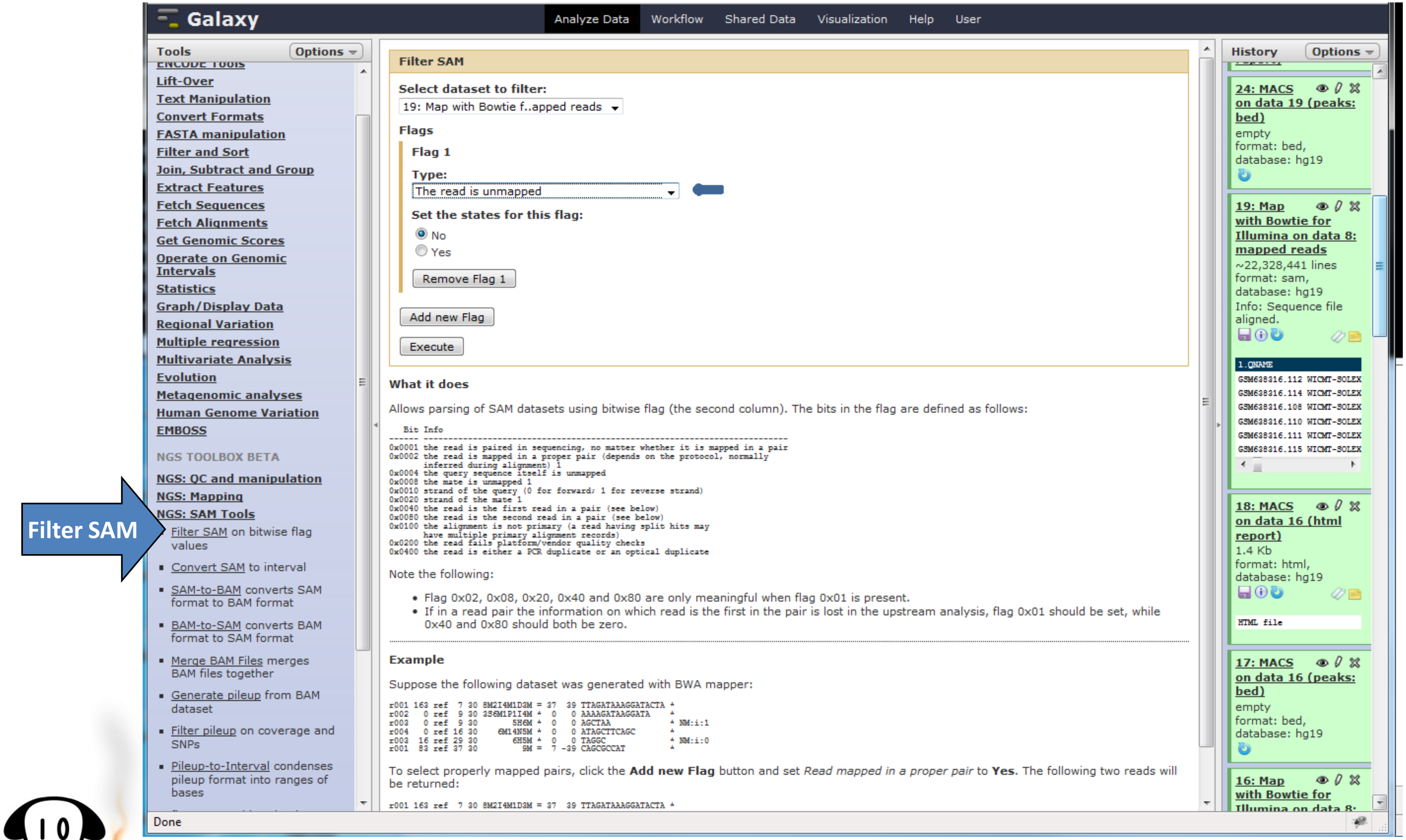
Suppress all alignments for a read if more than n reportable alignments exist (-m):
-1
-1 for no limit

Write all reads with a number of valid alignments exceeding the limit set with the -m option to a file (--max):
□

Write all reads that could not be aligned to a file (--un):
□

Whether or not to make Bowtie guarantee that reported singleton alignments are 'best' in terms of stratum and in terms of the quality values at the mismatched positions (--best):
Use best
Removes all strand bias. Only affects which alignments are reported by Bowtie. Runs slower with best option

Filtering unmapped reads



Filter SAM

Filter SAM

Select dataset to filter: 19: Map with Bowtie f..apped reads

Flags

Flag 1

Type: The read is unmapped

Set the states for this flag:

● No
○ Yes

What it does

Allows parsing of SAM datasets using bitwise flag (the second column). The bits in the flag are defined as follows:

```
   Bit Info
------ -----------------------------------------------------------------
0x0001 the read is paired in sequencing, no matter whether it is mapped in a pair
0x0002 the read is mapped in a proper pair (depends on the protocol, normally
       inferred during alignment) 1
0x0004 the query sequence itself is unmapped
0x0008 the mate is unmapped 1
0x0010 strand of the query (0 for forward; 1 for reverse strand)
0x0020 strand of the mate 1
0x0040 the read is the first read in a pair (see below)
0x0080 the read is the second read in a pair (see below)
0x0100 the alignment is not primary (a read having split hits may
       have multiple primary alignment records)
0x0200 the read fails platform/vendor quality checks
0x0400 the read is either a PCR duplicate or an optical duplicate
```

Note the following:

- Flag 0x02, 0x08, 0x20, 0x40 and 0x80 are only meaningful when flag 0x01 is present.
- If in a read pair the information on which read is the first in the pair is lost in the upstream analysis, flag 0x01 should be set, while 0x40 and 0x80 should both be zero.

Example

Suppose the following dataset was generated with BWA mapper:

```
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
r002   0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003   0 ref  9 30       5H6M *  0   0 AGCTAA            * NM:i:1
r004   0 ref 16 30    6M14N5M *  0   0 ATAGCTTCAGC       *
r003  16 ref 29 30       6H5M *  0   0 TAGGC             * NM:i:0
r001  83 ref 37 30         9M =  7 -39 CAGCGCCAT         *
```

To select properly mapped pairs, click the **Add new Flag** button and set *Read mapped in a proper pair* to **Yes**. The following two reads will be returned:

```
r001 163 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTA *
```

24: MACS on data 19 (peaks: bed)
empty
format: bed, database: hg19

19: Map with Bowtie for Illumina on data 8: mapped reads
~22,328,441 lines
format: sam, database: hg19
Info: Sequence file aligned.

```
1.QNAME
GSM638316.112 WICMT-SOLEX
GSM638316.114 WICMT-SOLEX
GSM638316.108 WICMT-SOLEX
GSM638316.110 WICMT-SOLEX
GSM638316.111 WICMT-SOLEX
GSM638316.115 WICMT-SOLEX
```

18: MACS on data 16 (html report)
1.4 Kb
format: html, database: hg19
HTML file

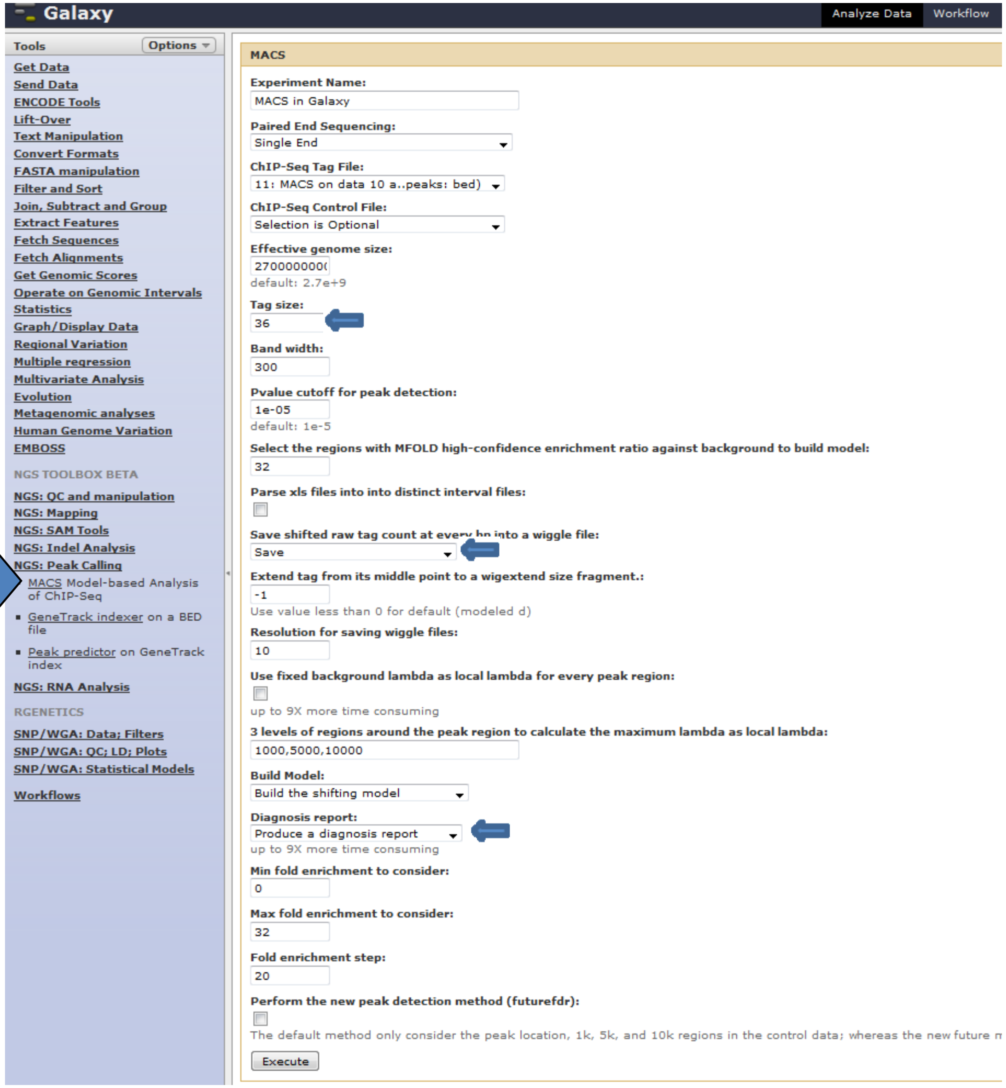
17: MACS on data 16 (peaks: bed)
empty
format: bed, database: hg19

16: Map with Bowtie for Illumina on data 8:

Filter SAM

Analysis of ChIP-seq experiments: MACS

MACS



MACS

Experiment Name: MACS in Galaxy

Paired End Sequencing: Single End

ChIP-Seq Tag File: 11: MACS on data 10 a..peaks: bed)

ChIP-Seq Control File: Selection is Optional

Effective genome size: 270000000(
default: 2.7e+9

Tag size: 36

Band width: 300

Pvalue cutoff for peak detection: 1e-05
default: 1e-5

Select the regions with MFOLD high-confidence enrichment ratio against background to build model: 32

Parse xls files into into distinct interval files: □

Save shifted raw tag count at every bp into a wiggle file: Save

Extend tag from its middle point to a wigextend size fragment.: -1
Use value less than 0 for default (modeled d)

Resolution for saving wiggle files: 10

Use fixed background lambda as local lambda for every peak region: □
up to 9X more time consuming

3 levels of regions around the peak region to calculate the maximum lambda as local lambda: 1000,5000,10000

Build Model: Build the shifting model

Diagnosis report: Produce a diagnosis report
up to 9X more time consuming

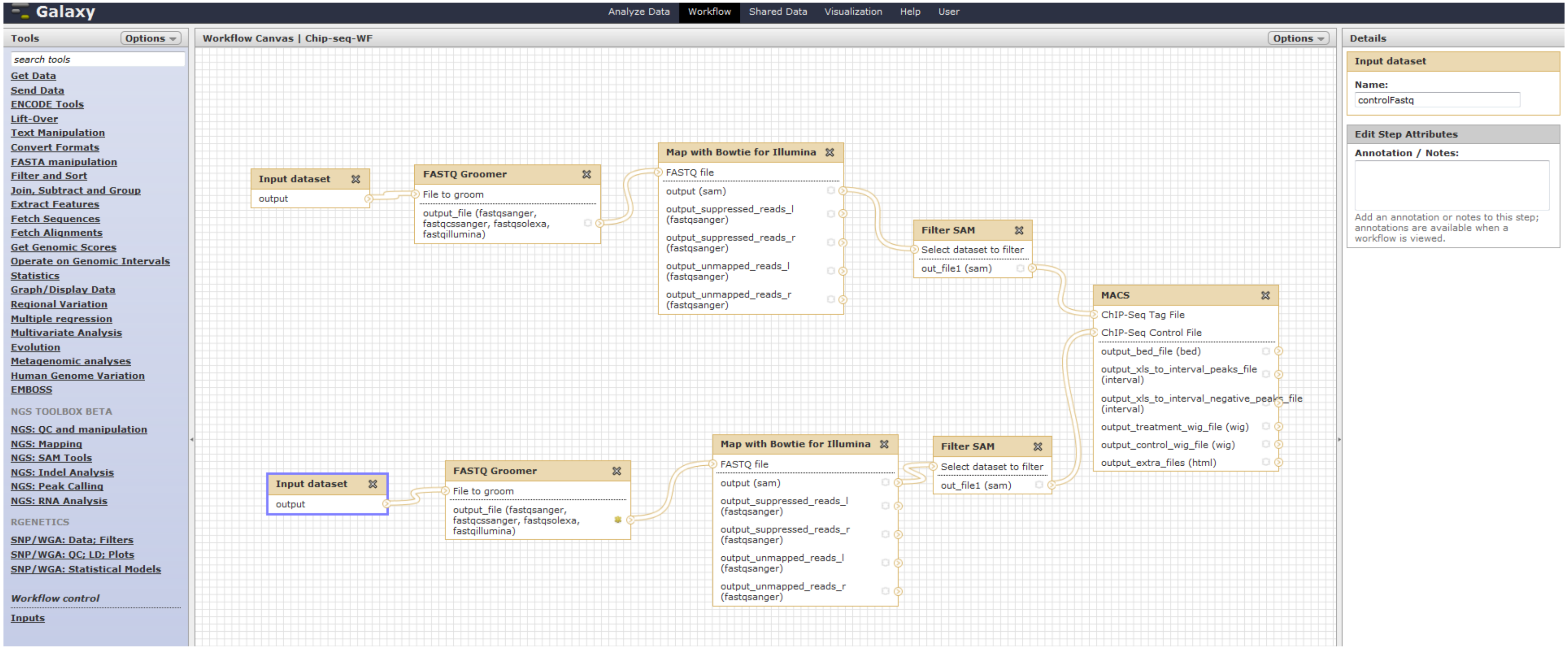
Min fold enrichment to consider: 0

Max fold enrichment to consider: 32

Fold enrichment step: 20

Perform the new peak detection method (futurefdr): □
The default method only consider the peak location, 1k, 5k, and 10k regions in the control data; whereas the new future m

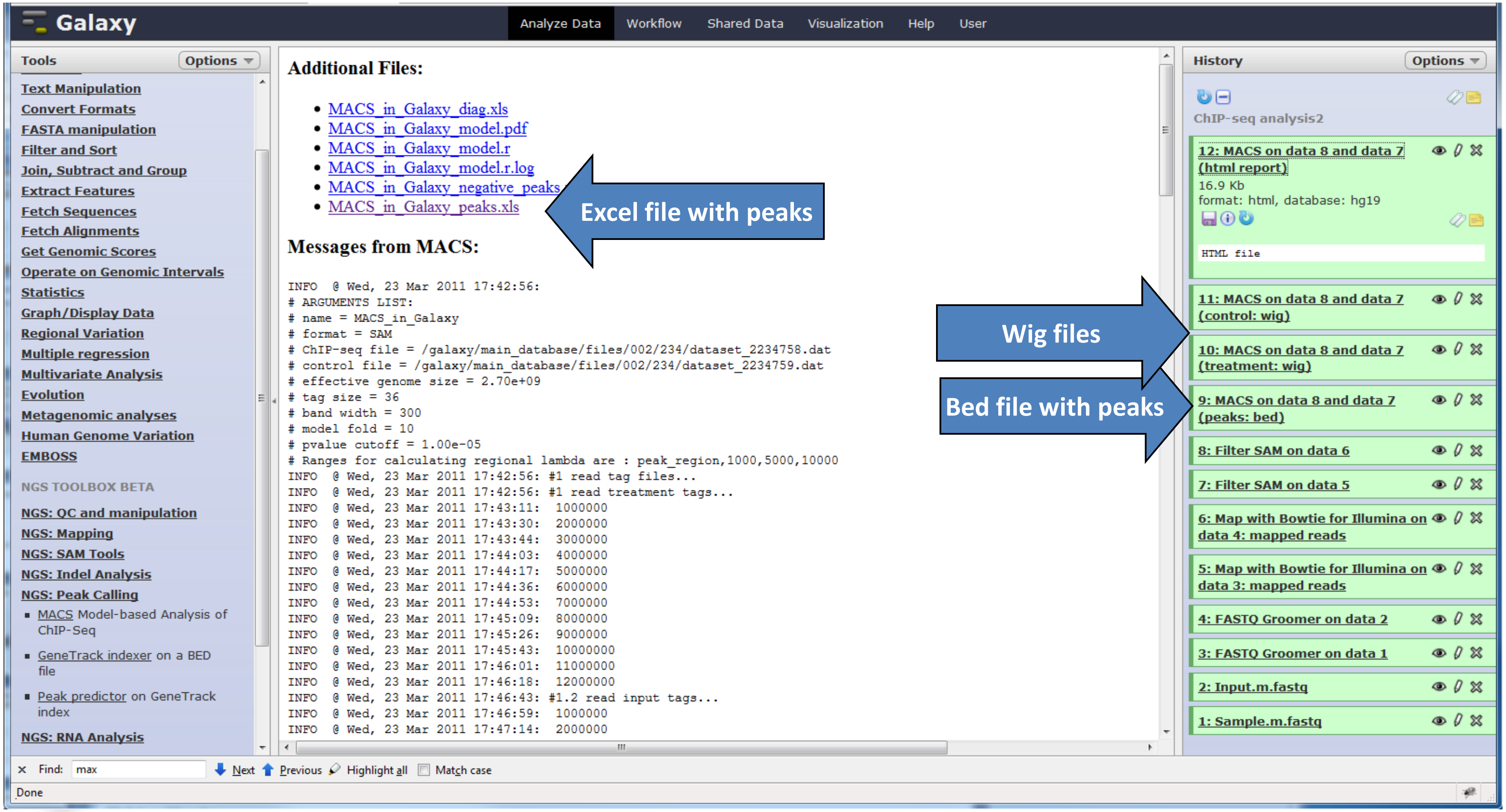
Workflow for ChIP-seq analysis



Name: controlFastq

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

MACS output



Additional Files:

- MACS_in_Galaxy_diag.xls
- MACS_in_Galaxy_model.pdf
- MACS_in_Galaxy_model.r
- MACS_in_Galaxy_model.r.log
- MACS_in_Galaxy_negative_peaks
- MACS_in_Galaxy_peaks.xls

Messages from MACS:

```
INFO  @ Wed, 23 Mar 2011 17:42:56:
# ARGUMENTS LIST:
# name = MACS_in_Galaxy
# format = SAM
# ChIP-seq file = /galaxy/main_database/files/002/234/dataset_2234758.dat
# control file = /galaxy/main_database/files/002/234/dataset_2234759.dat
# effective genome size = 2.70e+09
# tag size = 36
# band width = 300
# model fold = 10
# pvalue cutoff = 1.00e-05
# Ranges for calculating regional lambda are : peak_region,1000,5000,10000
INFO  @ Wed, 23 Mar 2011 17:42:56: #1 read tag files...
INFO  @ Wed, 23 Mar 2011 17:42:56: #1 read treatment tags...
INFO  @ Wed, 23 Mar 2011 17:43:11:  1000000
INFO  @ Wed, 23 Mar 2011 17:43:30:  2000000
INFO  @ Wed, 23 Mar 2011 17:43:44:  3000000
INFO  @ Wed, 23 Mar 2011 17:44:03:  4000000
INFO  @ Wed, 23 Mar 2011 17:44:17:  5000000
INFO  @ Wed, 23 Mar 2011 17:44:36:  6000000
INFO  @ Wed, 23 Mar 2011 17:44:53:  7000000
INFO  @ Wed, 23 Mar 2011 17:45:09:  8000000
INFO  @ Wed, 23 Mar 2011 17:45:26:  9000000
INFO  @ Wed, 23 Mar 2011 17:45:43:  10000000
INFO  @ Wed, 23 Mar 2011 17:46:01:  11000000
INFO  @ Wed, 23 Mar 2011 17:46:18:  12000000
INFO  @ Wed, 23 Mar 2011 17:46:43: #1.2 read input tags...
INFO  @ Wed, 23 Mar 2011 17:46:59:  1000000
INFO  @ Wed, 23 Mar 2011 17:47:14:  2000000
```

Excel file with peaks

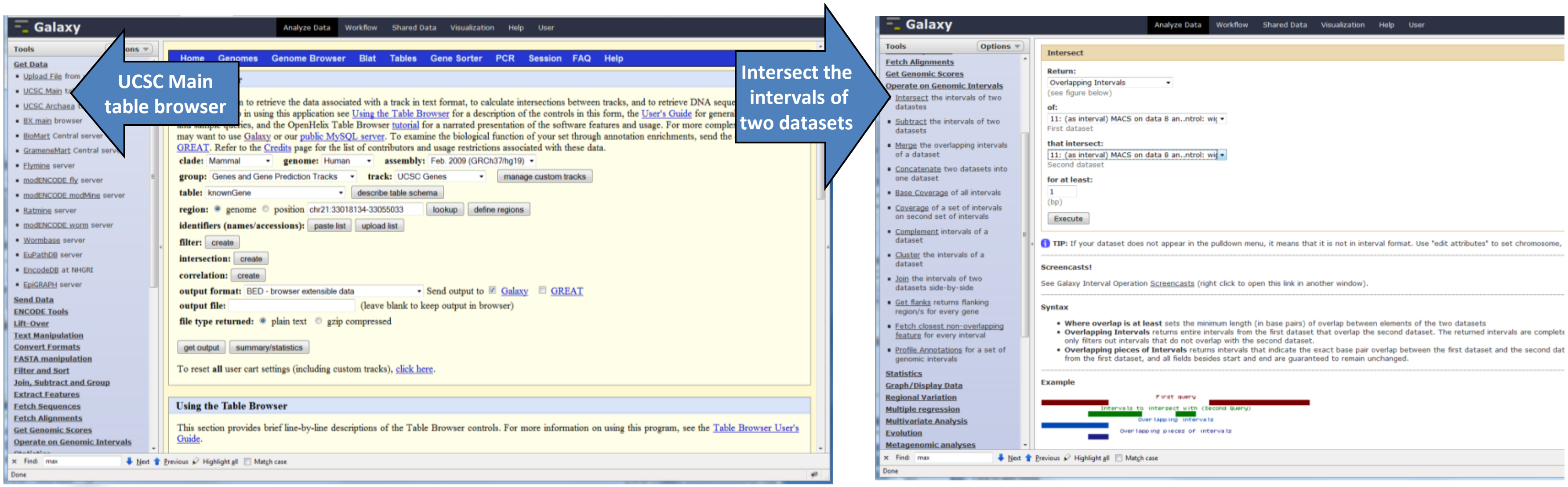
Wig files

Bed file with peaks

Analysis of ChIP-seq experiments: Intersect peaks with promoter regions

1. Download 1Kb regions upstream of genes from UCSC in bed format.
2. Get your bed file with peaks from MACS or other peak finding algorithm.
3. Intersect promoter bed file with peaks bed file.

(See Hot topics Dec 09, http://iona.wi.mit.edu/bio/education/hot_topics/galaxy/Galaxy.pdf)



UCSC Main table browser

n to retrieve the data associated with a track in text format, to calculate intersections between tracks, and to retrieve DNA seque
in using this application see Using the Table Browser for a description of the controls in this form, the User's Guide for genera
ies, and the OpenHelix Table Browser tutorial for a narrated presentation of the software features and usage. For more comple
may want to use Galaxy or our public MySQL server. To examine the biological function of your set through annotation enrichments, send the
GREAT. Refer to the Credits page for the list of contributors and usage restrictions associated with these data.

clade: Mammal **genome:** Human **assembly:** Feb. 2009 (GRCh37/hg19)

group: Genes and Gene Prediction Tracks **track:** UCSC Genes [manage custom tracks]

table: knownGene [describe table schema]

region: ● genome ○ position chr21:33018134-33055033 [lookup] [define regions]

identifiers (names/accessions): [paste list] [upload list]

filter: [create]

intersection: [create]

correlation: [create]

output format: BED - browser extensible data Send output to ☑ Galaxy ☐ GREAT

output file: (leave blank to keep output in browser)

file type returned: ● plain text ○ gzip compressed

[get output] [summary/statistics]

To reset **all** user cart settings (including custom tracks), click here.

Using the Table Browser

This section provides brief line-by-line descriptions of the Table Browser controls. For more information on using this program, see the Table Browser User's Guide.

Intersect the intervals of two datasets

Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Tools
Options
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Intersect the intervals of two datastes
Subtract the intervals of two datasets
Merge the overlapping intervals of a dataset
Concatenate two datasets into one dataset
Base Coverage of all intervals
Coverage of a set of intervals on second set of intervals
Complement intervals of a dataset
Cluster the intervals of a dataset
Join the intervals of two datasets side-by-side
Get flanks returns flanking region/s for every gene
Fetch closest non-overlapping feature for every interval
Profile Annotations for a set of genomic intervals
Statistics
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Metagenomic analyses
Find: max
Next
Previous
Highlight all
Match case
Done

Intersect

Return: Overlapping Intervals
(see figure below)

of: 11: (as interval) MACS on data 8 an..ntrol: wi
First dataset

that intersect: 11: (as interval) MACS on data 8 an..ntrol: wi
Second dataset

for at least: 1
(bp)

[Execute]

TIP: If your dataset does not appear in the pulldown menu, it means that it is not in interval format. Use "edit attributes" to set chromosome,

Screencasts!

See Galaxy Interval Operation Screencasts (right click to open this link in another window).

Syntax

- **Where overlap is at least** sets the minimum length (in base pairs) of overlap between elements of the two datasets
- **Overlapping Intervals** returns entire intervals from the first dataset that overlap the second dataset. The returned intervals are complete only filters out intervals that do not overlap with the second dataset.
- **Overlapping pieces of Intervals** returns intervals that indicate the exact base pair overlap between the first dataset and the second dat from the first dataset, and all fields besides start and end are guaranteed to remain unchanged.

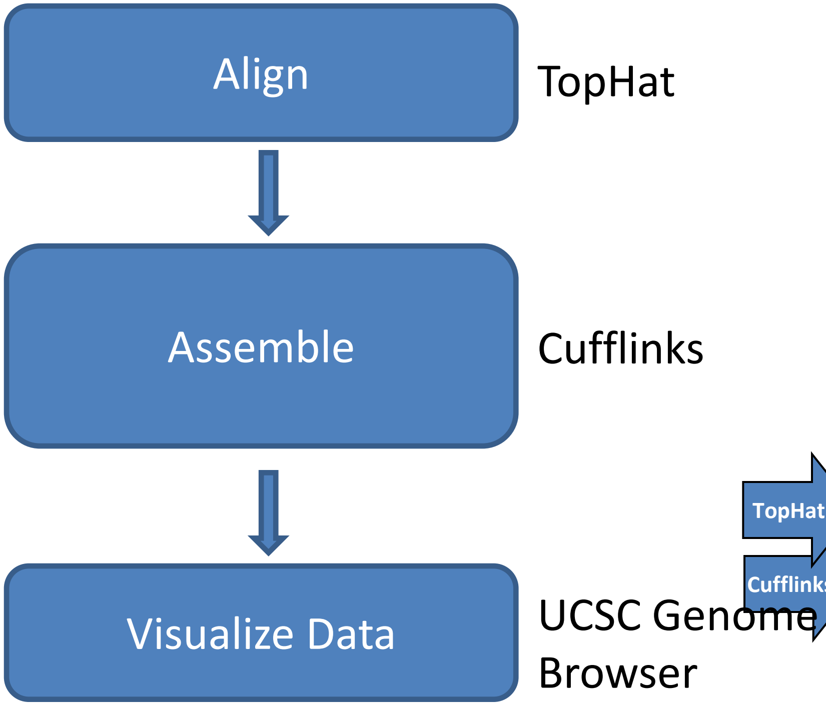
Example

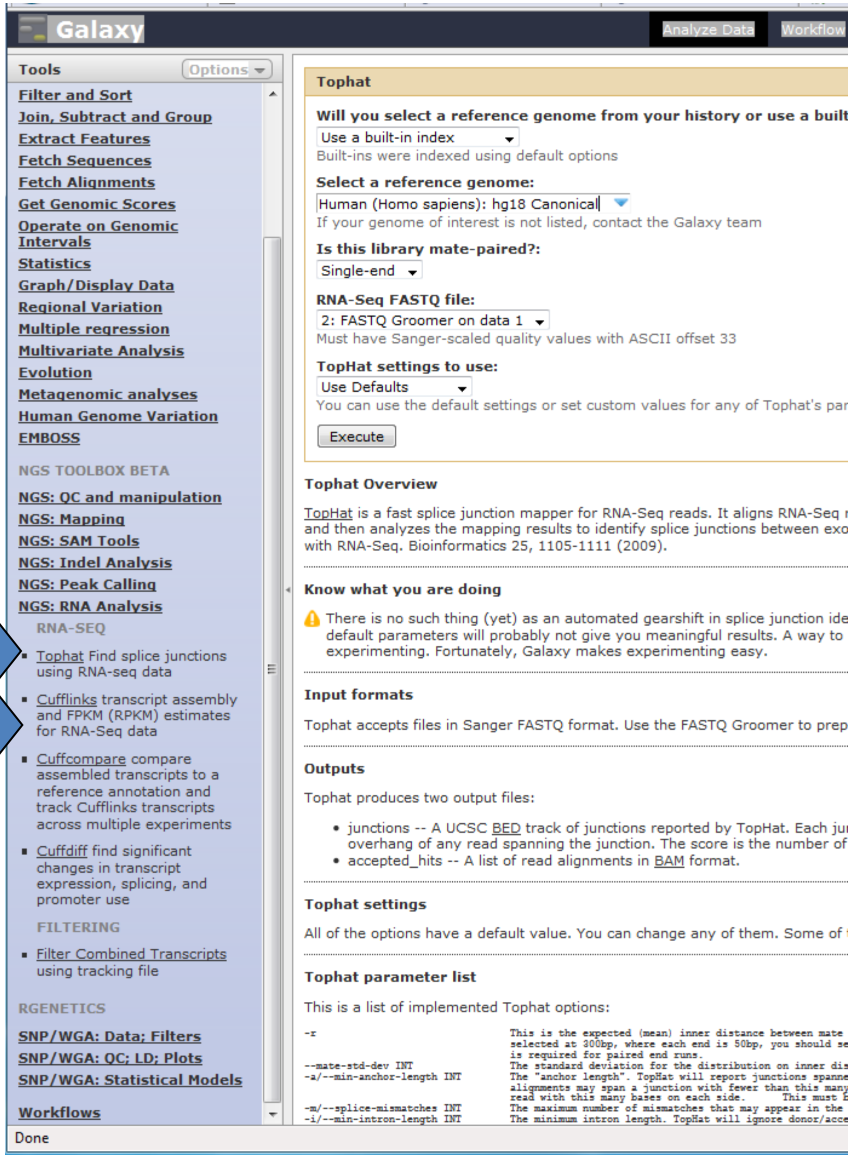
First query
Intervals to intersect with (Second Query)
Overlapping intervals
Overlapping pieces of intervals

Talk Outline

- Introduction to Galaxy
- Data upload
- Format conversion and quality control tools
- Mapping
- Analysis of ChIP-seq experiments with MACs
- Analysis of RNA-seq experiments with Tuxedo tools
- Demo

Expression Profiling Workflow


Align
TopHat
Assemble
Cufflinks
Visualize Data
UCSC Genome Browser



Tools Options

- Filter and Sort
- Join, Subtract and Group
- Extract Features
- Fetch Sequences
- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Metagenomic analyses
- Human Genome Variation
- EMBOSS

NGS TOOLBOX BETA

- NGS: QC and manipulation
- NGS: Mapping
- NGS: SAM Tools
- NGS: Indel Analysis
- NGS: Peak Calling
- NGS: RNA Analysis

RNA-SEQ

- Tophat Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use

FILTERING

- Filter Combined Transcripts using tracking file

RGENETICS

- SNP/WGA: Data; Filters
- SNP/WGA: QC; LD; Plots
- SNP/WGA: Statistical Models

Workflows

Tophat

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

Select a reference genome:
Human (Homo sapiens): hg18 Canonical
If your genome of interest is not listed, contact the Galaxy team

Is this library mate-paired?:
Single-end

RNA-Seq FASTQ file:
2: FASTQ Groomer on data 1
Must have Sanger-scaled quality values with ASCII offset 33

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

Execute

Tophat Overview

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq r
and then analyzes the mapping results to identify splice junctions between exo
with RNA-Seq. Bioinformatics 25, 1105-1111 (2009).

Know what you are doing

There is no such thing (yet) as an automated gearshift in splice junction ide
default parameters will probably not give you meaningful results. A way to
experimenting. Fortunately, Galaxy makes experimenting easy.

Input formats

Tophat accepts files in Sanger FASTQ format. Use the FASTQ Groomer to prep

Outputs

Tophat produces two output files:

- junctions -- A UCSC BED track of junctions reported by TopHat. Each jun
  overhang of any read spanning the junction. The score is the number of
- accepted_hits -- A list of read alignments in BAM format.

Tophat settings

All of the options have a default value. You can change any of them. Some of

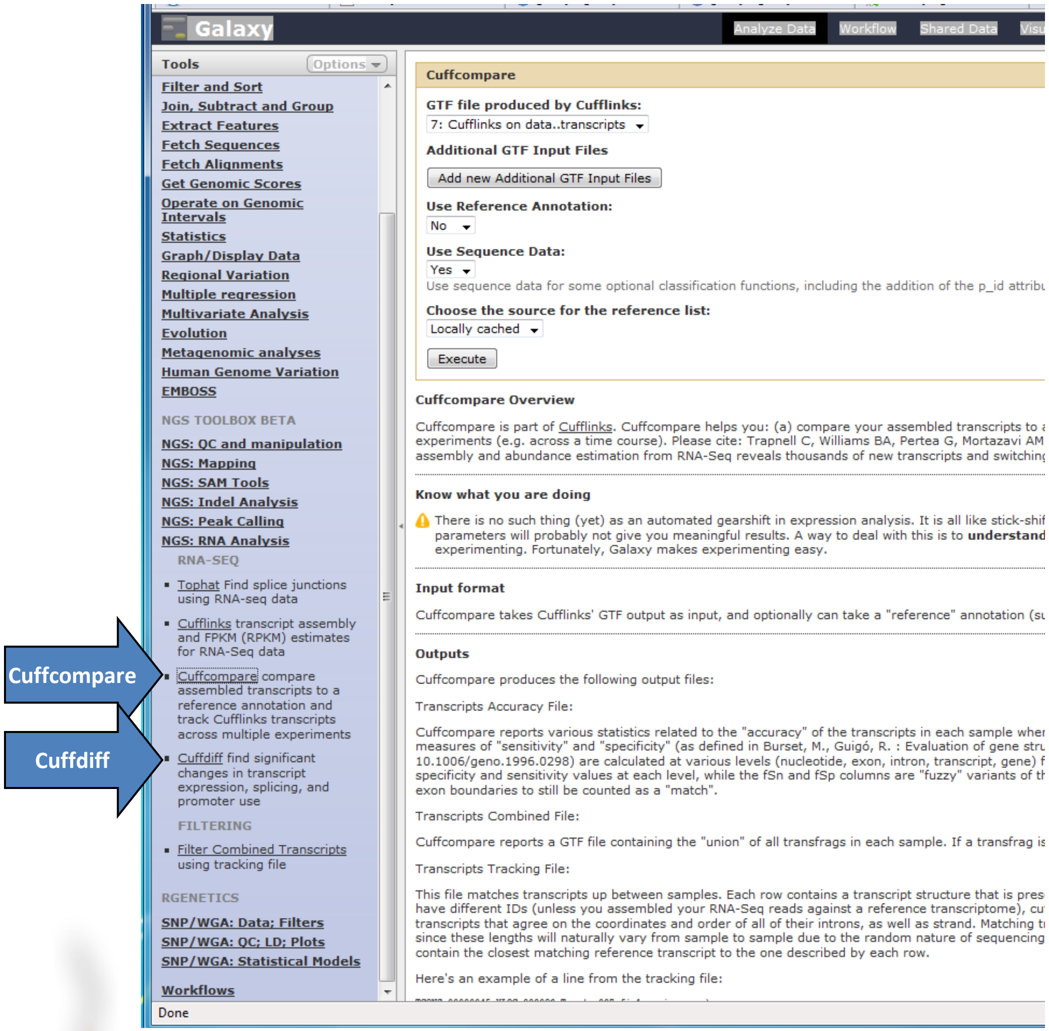
Tophat parameter list

This is a list of implemented Tophat options:

```
-r                           This is the expected (mean) inner distance between mate
                             selected at 300bp, where each end is 50bp, you should se
                             is required for paired end runs.
--mate-std-dev INT           The standard deviation for the distribution on inner dis
-a/--min-anchor-length INT   The "anchor length". TopHat will report junctions spanne
                             alignments may span a junction with fewer than this many
                             read with this many bases on each side.      This must b
-m/--splice-mismatches INT   The maximum number of mismatches that may appear in the
-i/--min-intron-length INT   The minimum intron length. TopHat will ignore donor/acce
```

Done

Other tools for expression profiling



- Tophat Find splice junctions using RNA-seq data
- Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
- Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff find significant changes in transcript expression, splicing, and promoter use
- Filter Combined Transcripts using tracking file

Cuffcompare

GTF file produced by Cufflinks: 7: Cufflinks on data..transcripts

Additional GTF Input Files

Add new Additional GTF Input Files

Use Reference Annotation: No

Use Sequence Data: Yes

Use sequence data for some optional classification functions, including the addition of the p_id attribu

Choose the source for the reference list: Locally cached

Execute

Cuffcompare Overview

Cuffcompare is part of Cufflinks. Cuffcompare helps you: (a) compare your assembled transcripts to a
experiments (e.g. across a time course). Please cite: Trapnell C, Williams BA, Pertea G, Mortazavi AM
assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching

Know what you are doing

⚠ There is no such thing (yet) as an automated gearshift in expression analysis. It is all like stick-shif
parameters will probably not give you meaningful results. A way to deal with this is to **understand**
experimenting. Fortunately, Galaxy makes experimenting easy.

Input format

Cuffcompare takes Cufflinks' GTF output as input, and optionally can take a "reference" annotation (su

Outputs

Cuffcompare produces the following output files:

Transcripts Accuracy File:

Cuffcompare reports various statistics related to the "accuracy" of the transcripts in each sample wher
measures of "sensitivity" and "specificity" (as defined in Burset, M., Guigó, R. : Evaluation of gene stru
10.1006/geno.1996.0298) are calculated at various levels (nucleotide, exon, intron, transcript, gene) f
specificity and sensitivity values at each level, while the fSn and fSp columns are "fuzzy" variants of th
exon boundaries to still be counted as a "match".

Transcripts Combined File:

Cuffcompare reports a GTF file containing the "union" of all transfrags in each sample. If a transfrag is

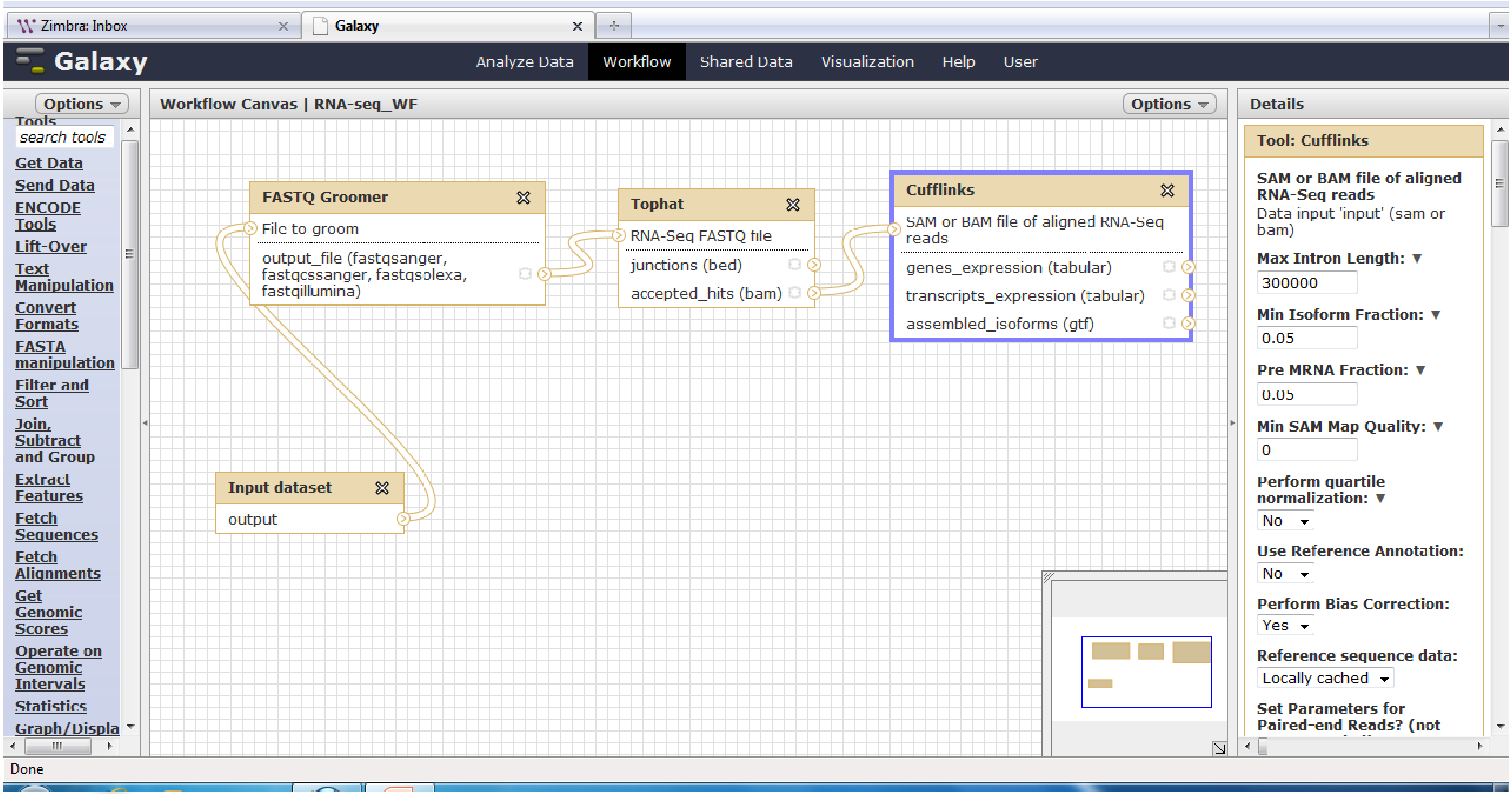
Transcripts Tracking File:

This file matches transcripts up between samples. Each row contains a transcript structure that is pres
have different IDs (unless you assembled your RNA-Seq reads against a reference transcriptome), cu
transcripts that agree on the coordinates and order of all of their introns, as well as strand. Matching tr
since these lengths will naturally vary from sample to sample due to the random nature of sequencing
contain the closest matching reference transcript to the one described by each row.

Here's an example of a line from the tracking file:

- Cuffcompare: compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
- Cuffdiff: find significant changes in transcript expression, splicing, and promoter use

Workflow for RNA-seq analysis



Tool: Cufflinks

SAM or BAM file of aligned RNA-Seq reads
Data input 'input' (sam or bam)

Max Intron Length: ▼
300000

Min Isoform Fraction: ▼
0.05

Pre MRNA Fraction: ▼
0.05

Min SAM Map Quality: ▼
0

Perform quartile normalization: ▼
No

Use Reference Annotation:
No

Perform Bias Correction:
Yes

Reference sequence data:
Locally cached

Set Parameters for Paired-end Reads? (not

Workflow/Demo for ChIP-seq analysis

1. Workflow for quality control
2. Workflow for mapping and running MACS
3. Workflow for RNA-seq