



RNA-Seq Data Analysis using Galaxy

RNASeq Data Analysis Bootcamp Week4

Alper Kucukural
BioCore
03/04/2014

RNA-Seq Data Analysis using Galaxy

Alper Kucukural

BioCore

03/04/2014



What is Galaxy?

Galaxy is

- an integrated tool management system with a user-friendly (GUI).
- It is designed for running multiple bioinformatics tools.

Galaxy Basics

- Data upload.
- Grooming the data.
- FastQC Quality Check
- Filtering FastQ files

Galaxy RNASeq Tools

- Read alignments (Tophat2)
- UCSC Genome Browser
- Coverage Calculation
- BedGraph to bigWig

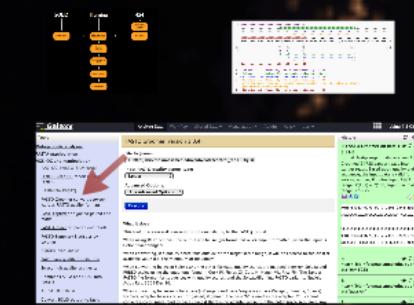
Galaxy Basics

Galaxy Overview



Grooming the Data

When your read qualities are not formatted for galaxy...

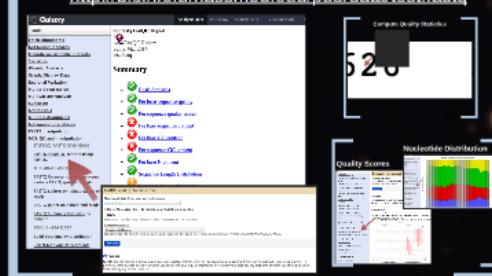


Data Upload



FASTQC and other QC

Upload this file to your galaxy; <http://bioinfo.umassmed.edu/pub/data/test.fastq>



Galaxy Overview

galaxy.umassmed.edu or
usegalaxy.org

The screenshot shows the Galaxy web interface at galaxy.umassmed.edu. A red arrow points to the left menu, which contains various genomic analysis tools. Another red arrow points to the 'History' panel on the right, which lists recent data operations.

Left menu: Tools, Get Data, Send Data, ENCODE Tools, Lift-Over, Text Manipulation, Filter and Sort, Join, Subtract and Group, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Wavelet Analysis, Graph/Display Data, Regional Variation, Multiple regression, Multivariate Analysis, Evolution, Motif Tools, Multiple Alignments.

History:

- Unnamed history (184.6 MB)
 - 62: Join two Datasets on data 61 and data 48
 - 61: Convert on data 47
 - 59: Join two Datasets on data 47 and data 48
 - 48: Cut on data 29
 - 33,063 lines
 - format: tabular, database: mm10
 - 1 2
 - Xkr4 NM_001011874
 - Rp1 NM_001195662
 - Rp1 NM_011283
 - Sox17 NM_001289464
 - Sox17 NM_001289465
 - Sox17 NM_001289467
 - 47: Group on data 45
 - 52 lines

Data Upload

The screenshot shows the Galaxy web interface with a dark background and orange highlights. The top navigation bar includes 'Galaxy', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', and 'Admin'. On the left, a sidebar titled 'Tools' contains a search bar and a list of links under 'Get Data': 'Upload File from your computer', 'UCSC Main table browser', 'UCSC Test table browser', 'UCSC Archaea table browser', 'BX table browser', 'EBI SRA ENA SRA', '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', and 'modENCODE modMine server'. A red arrow points to the 'Upload File from your computer' link. The main content area is titled 'Upload File (version 1.1.3)'. It has sections for 'File Format' (set to 'Auto-detect'), 'File' (with a 'Choose File' button and a note about file size limits), and 'URL/Text' (containing two URLs: 'http://bioinfo.umassmed.edu/pub/data/control_rep1.1.fq' and 'http://bioinfo.umassmed.edu/pub/data/control_rep1.2.fq'). A red arrow points to the URL input field. Below it is a note: 'Here you may specify a list of URLs (one per line) or paste the contents of a file'. The 'Files uploaded via FTP' section shows a table with columns 'File', 'Size', and 'Date', stating 'Your FTP upload directory contains no files.' At the bottom, there's a note about using the Galaxy server for FTP uploads and a 'Convert spaces to tabs:' link.

Galaxy

Analyze Data Workflow Shared Data Visualization Admin

Tools

search tools

Get Data

Upload File from your computer

UCSC Main table browser

UCSC Test table browser

UCSC Archaea table browser

BX table browser

EBI SRA ENA SRA

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

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 files, use the URL method (below) or FTP (if enabled by the site administrator)

URL/Text:

http://bioinfo.umassmed.edu/pub/data/control_rep1.1.fq
http://bioinfo.umassmed.edu/pub/data/control_rep1.2.fq

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

Files uploaded via FTP:

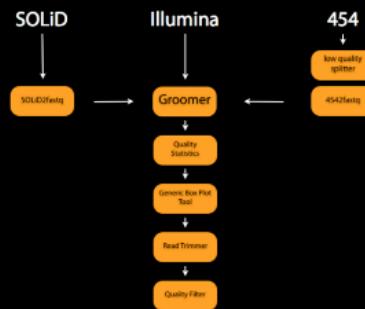
File	Size	Date
Your FTP upload directory contains no files.		

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

Convert spaces to tabs:

Grooming the Data

When your read qualities are not formatted for galaxy...



Galaxy

Analyze Data Workflow Shared Data Visualization Admin Help User

Tools

Metagenomic analyses
FASTA manipulation
NGS: QC and manipulation
FASTQC: FASTQ/SAM/BAM
FastQC: Read QC reports using FastQC
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
AB-SOLID DATA
Convert SOLiD output to fastq

FASTQ Groomer (version 1.0.4)

File to groom:
1: http://bioinfo.umassmed.edu/pub/data/control_rep1.1.fq

Input FASTQ quality scores type:
Sanger

Advanced Options:
Hide Advanced Options

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

History

5: FASTQ Groomer on data 1

3.1 MB
format: fastqsanger, database: mm10
Groomed 24788 sanger reads into sanger reads. Based upon quality and sequence, the input data is valid for: solexa, sanger, illumina Input ASCII range: 'I'(105) – 'I'(105) Input decimal range: 72 – 72

HWI-ST333_0273_FC:7:1205:9843:15143
CAAGGAACAGCATGACCAGCAGAAAATACCCAGT
+
|||||||||||||||||||||||||||||||||||||||||||
HWI-ST333_0273_FC:7:2116:8217:37625
GTCAGCTCCGTATTTCTCCAGGCCACTGTACACTA

4:
<http://bioinfo.umassmed.edu/pub/data/mm10.fa>

3:
<http://bioinfo.umassmed.edu/pub/data/ucsc.gtf>

2:

SOLID

Illumina

454





```

0.2.....26...31.....41
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)

```

Galaxy

[Analyze Data](#)
[Workflow](#)
[Shared Data ▾](#)
[Visualization ▾](#)
[Admin](#)
[Help ▾](#)
[User ▾](#)


Using 7.2 GB

Tools

Metagenomic analyses

FASTA manipulation

NGS: QC and manipulation

[FASTQC: FASTQ/SAM/BAM](#)
[FastQC:Read QC reports using FastQC](#)
[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](#)
[AB-SOLID DATA](#)
[Convert SOLID output to fastq](#)

FASTQ Groomer (version 1.0.4)

File to groom:

Input FASTQ quality scores type:

Advanced Options:

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

History

5: FASTQ Groomer on data 1

3.1 MB

format: fastqsanger, database: mm10
Groomed 24788 sanger reads into sanger reads. Based upon quality and sequence, the input data is valid for: solexa, sanger, illumina Input ASCII range: 'I'(105) – 'I'(105) Input decimal range: 72 – 72

@HWI-ST333_0273_FC:7:1205:9843:15143:

CAAGGAAGCACATGACCGAGCAGAAATACCCAGTT

+

ii

@HWI-ST333_0273_FC:7:2116:8217:37625:

GTCAGCTCCTGATGTTCTCCAGGCCACTGTACACT,

4:

<http://bioinfo.umassmed.edu/pub/data/mm10.fa>

3:

<http://bioinfo.umassmed.edu/pub/data/ucsc.gtf>

2:

FASTQC and other QC

Upload this file to your galaxy;
<http://bioinfo.umassmed.edu/pub/data/test.fasta>

Galaxy

Analyze Data Workflow Shared Data Visualization

Tools

- Fetch Alignments
- Get Genomic Scores
- Operate on Genomic Intervals
- Statistics
- Wavelet Analysis
- Graph/Display Data
- Regional Variation
- Multiple regression
- Multivariate Analysis
- Evolution
- Motif Tools
- Multiple Alignments
- Metagenomic analyses
- FASTA manipulation
- NGS: QC and manipulation

FASTQC: FASTQ/SAM/BAM

FastQC:Read QC reports using FastQC

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

FastQC:Read QC (version 0.52)

Short read data from your current history:
3: test.fasta

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

Letters and numbers only please – other characters will be removed

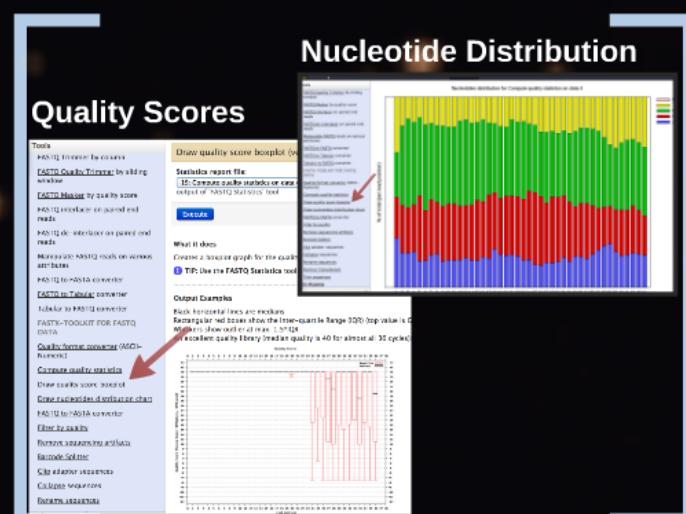
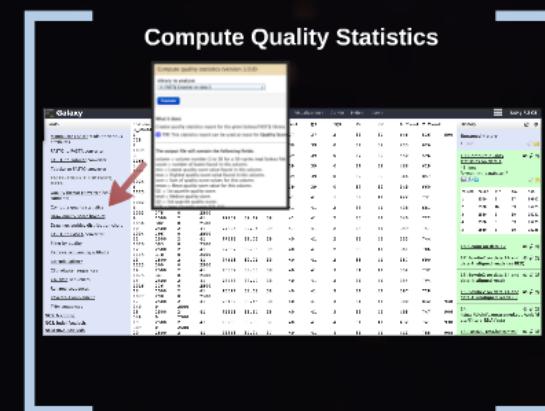
Contaminant list:
Selection is Optional

tab delimited file with 2 columns: name and sequence. For example: Illumina Small RNA RT Primer CAAGCAGAAAGACGGCATACGA

Execute

Purpose

FastQC aims to provide a simple way to do some quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a modular set of analyses which you can use to give a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.



Compute Quality Statistics

Tools	column G_Count
Manipulate FASTQ reads on various attributes	1 728
FASTQ to FASTA converter	2 1039
FASTQ to Tabular converter	3 1166
Tabular to FASTQ converter	4 1081
FASTX-TOOLKIT FOR FASTQ DATA	5 1025
Quality format converter (ASCII-Numeric)	6 1125
Compute quality statistics	7 1002
	8 1003
Draw quality score boxplot	9 2500
Draw nucleotides distribution chart	10 1056
FASTQ to FASTA converter	11 1004
Filter by quality	12 1010
Remove sequencing artifacts	13 1115
Barcode Splitter	14 1123
Clip adapter sequences	15 1075
Collapse sequences	16 1016
Rename sequences	17 1077
Reverse-Complement	18 340
Trim sequences	19 349
NGS: Mapping	20 349
NGS: Indel Analysis	389
NGS: RNA Analysis	20 2500

Compute quality statistics (version 1.0.0)

Library to analyse:
4: FASTQ Groomer on data 3

Execute

What it does

Creates quality statistics report for the given Solexa/FASTQ library.

TIP: This statistics report can be used as input for Quality Score Boxplot.

The output file will contain the following fields:

column = column number (1 to 36 for a 36-cycles read Solexa file).
count = number of bases found in this column.
min = Lowest quality score value found in this column.
max = Highest quality score value found in this column.
sum = Sum of quality score values for this column.
mean = Mean quality score value for this column.
Q1 = 1st quartile quality score.
med = Median quality score.
Q3 = 3rd quartile quality score.
IQR = Inter Quartile Range (Q3 - Q1)

	med	Q3	IQR	1W	rW	A_Count	C_Count		History
1	37	37	2	32	37	646	536	590	Unnamed history
2	39	39	0	39	39	453	626		1.0 GB
3	39	39	0	39	39	450	596		19: Compute quality statistics on data 4
4	39	39	0	39	39	466	619		45 lines
5	39	39	0	39	39	346	857		format: txt, database: ?
6	39	39	0	39	39	349	660		
7	41	41	1	39	41	422	731		
8	41	41	2	36	41	438	681		
9	41	41	2	36	41	340	722		18: Group on data 17
10	41	41	2	36	41	359	741		17: Bowtie2 on data 14 and data 4: aligned reads (as BED)
11	41	41	2	36	41	353	744		16: Bowtie2 on data 14 and data 4: aligned reads
12	41	41	2	36	41	361	706		15: Bowtie2 on data 14 and data 4: unaligned reads (L)
13	41	41	2	36	41	387	690		14: http://bioinfo.umassmed.edu/pub/datasets/filter/rRNA.fasta
14	41	41	3	34	41	352	732		
15	41	41	3	34	41	354	794		
16	41	41	3	34	41	397	728		
17	41	41	3	34	41	380	832	948	
18	41	41	3	34	41	498	747	906	
19	41	41	4	31	41	432	741	938	
20	41	41	4	31	41	413	785	908	13: FastQC test.fastq.html

Nucleotide Distribution

Quality Scores

Tools

- [FASTQ Trimmer by column](#)
- [FASTQ Quality Trimmer by sliding window](#)
- [FASTQ Masker by quality score](#)
- [FASTQ interlacer on paired end reads](#)
- [FASTQ de-interlacer on paired end reads](#)
- [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](#)

Draw quality score boxplot (v)

Statistics report file:

19: Compute quality statistics on data 4
output of 'FASTQ Statistics' tool

Execute

What it does

Creates a boxplot graph for the quality

TIP: Use the **FASTQ Statistics** tool

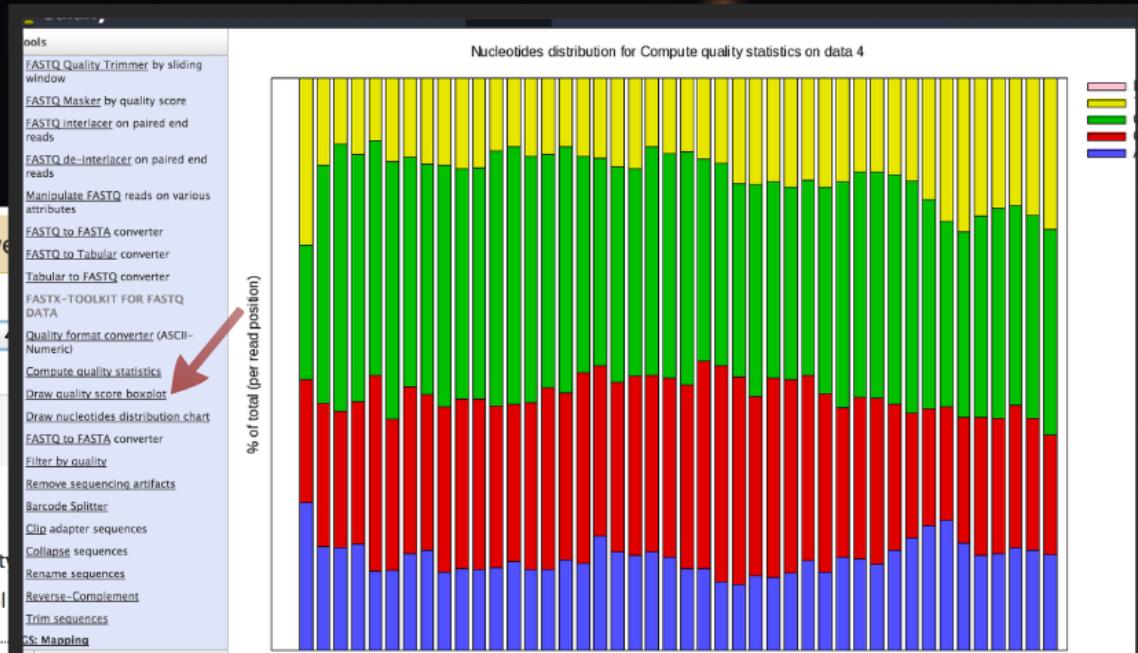
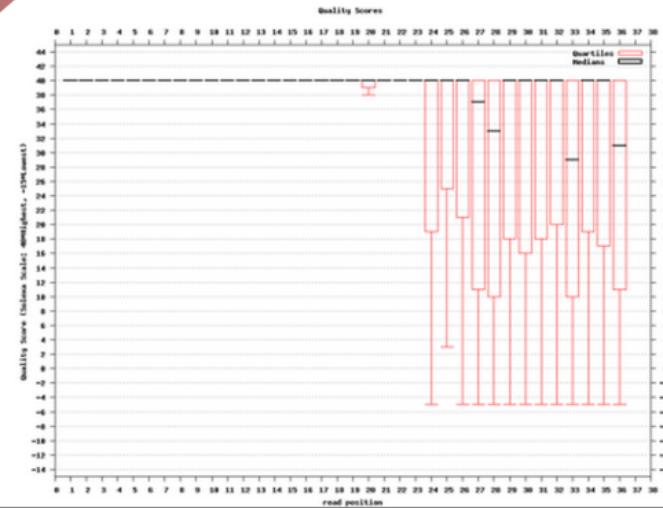
Output Examples

Black horizontal lines are medians

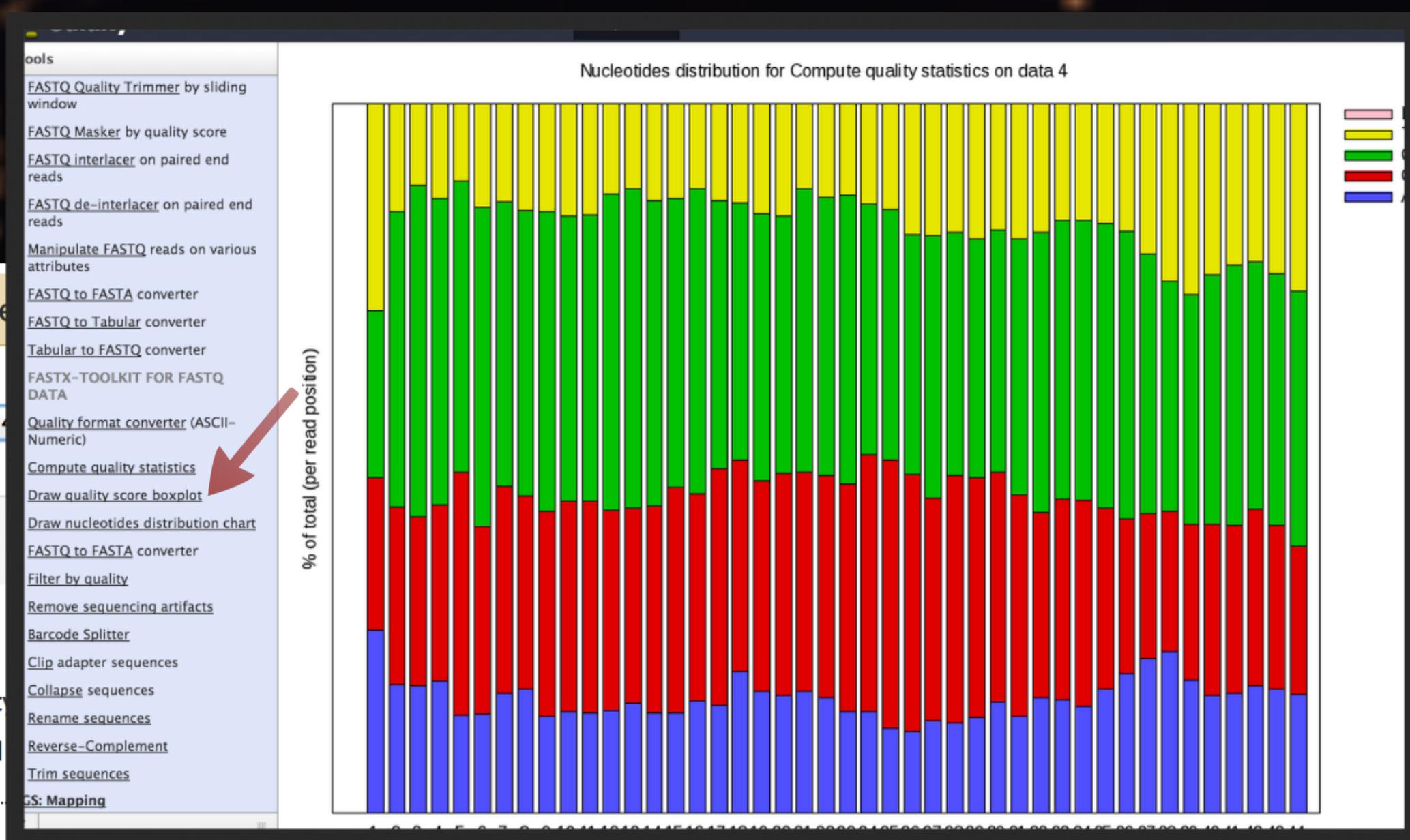
Rectangular red boxes show the Inter-quartile Range (IQR) (top value is Q3)

Whiskers show outlier at max. $1.5 \times \text{IQR}$

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



Nucleotide Distribution



Read Trimming

Galaxy

Analyze Data

Tools

- 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**
- FASTQ interlacer on paired end reads**
- FASTQ de-interlacer on paired end reads**
- Manipulate FASTQ reads on various attributes**
- FASTQ to FASTA converter**
- FASTQ to Tabular converter**
- Tabular to FASTQ converter**
- FASTX-TOOLKIT FOR FASTQ DATA**

FASTQ Trimmer (version 1.0.0)

FASTQ File:
4: FASTQ Groomer on data 3

Define Base Offsets as:
 Absolute Values
Use Absolute for fixed length reads (Illumina, SOLiD)
Use Percentage for variable length reads (Roche/454)

Offset from 5' end:

Values start at 0, increasing from the left

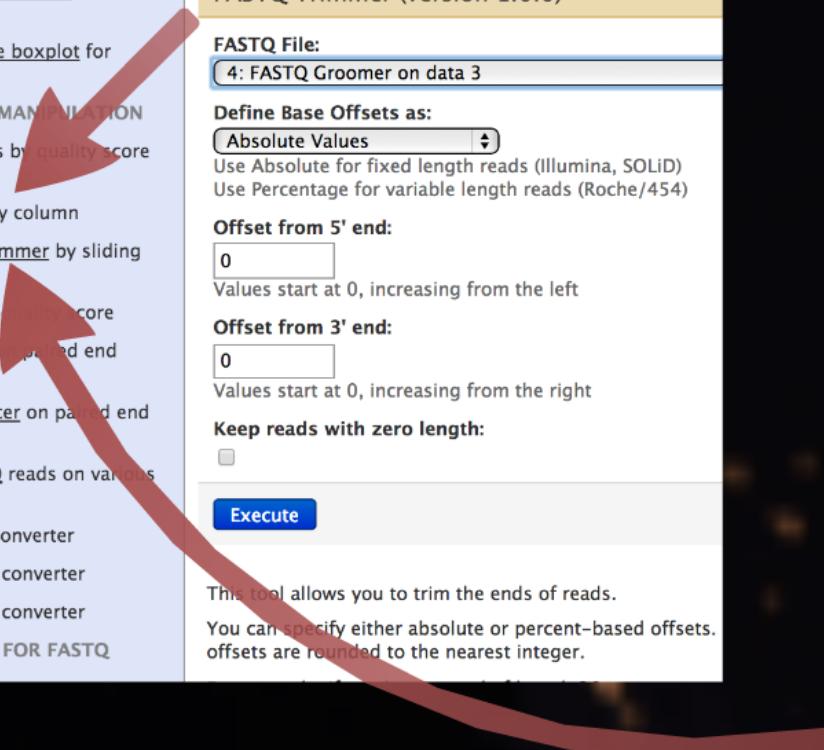
Offset from 3' end:

Values start at 0, increasing from the right

Keep reads with zero length:

Execute

This tool allows you to trim the ends of reads.
You can specify either absolute or percent-based offsets.
offsets are rounded to the nearest integer.



Quality Filtering

FASTQ Quality Trimmer (version 1.0.0)

FASTQ File:
15: Bowtie2 on data 14 and data 4: unaligned reads (L)

Keep reads with zero length:

Trim ends:
 5' and 3'
 3' only
 5' only

Window size:

Step Size:

Maximum number of bases to exclude from the window during aggregation:

Aggregate action for window:
 min score
 max score

Trim until aggregate score is:
 >=
 <=

Quality Score:

Execute

Example: Filtering Ribosomal Reads and Counting

- Upload Ribosomal RNA sequences.

<http://bioinfo.umassmed.edu/pub/data/filter/rRNA.fasta>

History  

14:   
<http://bioinfo.umassmed.edu/pub/database/filter/rRNA.fasta>

4 sequences
format: fasta, database: hg19
uploaded fasta file

Mapping to rRNAs

Tools

NGS: Mapping

Map with Bowtie for Illumina

Map with BWA (Version 0.7.5) This new version BWA (0.7.5) use 'mem' algorithm for mapping, doesn't need 'aln', 'samse', 'sampe' and picard AddOrReplace anymore.

Bowtie2 is a short-read aligner

Map with BFAST

Megablast compare short reads against htgs, nt, and wgs databases

Parse blast XML output

Map with PerM for SOLiD and Illumina

Re-align with SRMA

Map with Mosaik

NGS: Indel Analysis

NGS: RNA Analysis

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: Peak Calling

Bowtie2 (version 0.2)

Is this library mate-paired?

FASTQ file:

4: FASTQ Groomer on data 3

Nucleotide-space: Must have Sanger-scaled quality values with ASCII

Write unaligned reads to separate file(s):

Will you select a reference genome from your history or use a built-in?

Use one from the history

Built-ins were indexed using default options

Select the reference genome:

14: http://bioinfo.umassmed.edu/pub/data/filter/rRNA.fasta

Specify the read group for this file:

No

Parameter Settings:

Use defaults

You can use the default

Execute

Convert from BAM to BED (version 0.1.0)

Convert the following BAM file to BED:
[BAM file] on data 14 and data 3: standard fastq

What type of BED output would you like:
[Create 4-column BED file]

Import aligned BAM alignments as separate BED entries

Use alignment's edit-distance for BED name

Use other **NAME** BAM alignment tag as the BED name

Submit

Group the Data and Count

The screenshot shows the Galaxy 'Group' tool interface (version 2.0.0). The left sidebar lists various tools: Tools, Join, Subtract and Group, Join two Datasets, Compare two Datasets, Subtract Whole Dataset, Group by column, Group data by a column and perform aggregate operation on other columns, Column Join, Convert Formats, Extract Features, Fetch Sequences, Fetch Alignments, Get Genomic Scores, Operate on Genomic Intervals, Statistics, Wavelet Analysis, Graph/Display Data, and Regional Variation. A red arrow points from the 'Group data by a column and perform aggregate operation on other columns.' link to the 'Column Join' section in the main panel. Another red arrow points from the 'Count' dropdown in the 'Operation 1' section to the 'On column:' dropdown.

Group (version 2.0.0)

Select data:
17: Bowtie2 on data 14 and data 4: aligned reads (as BED)
Dataset missing? See TIP below.

Group by column:
c1

Ignore case while grouping?:

Operations

Operation 1

Type:
Count

On column:
c1

Round result to nearest integer?:
NO

Analyze Data Workflow Shared Data Visualization Admin Help User

3000072055985-55#tRNA		10
gi 124517659 ref NR_003286.1 _Homo_sapiens_18S_ribosomal_RNA_(LOC100008588)		40
gi 124517661 ref NR_003287.1 _Homo_sapiens_28S_ribosomal_RNA_(LOC100008589)		1353
gi 1243272596 ref NR_003285.2 _Homo_sapiens_5.8S_ribosomal_RNA_(LOC100008587)		32

Mapping to rRNAs

Tools

NGS: Mapping

[Map with Bowtie for Illumina](#)

[Map with BWA \(Version 0.7.5\)](#) This new version BWA (0.7.5) use 'mem' algorithm for mapping, doesn't need 'aln', 'samse', 'sampe' and picard AddOrReplace anymore.

[Bowtie2 is a short-read aligner](#)

[Map with BFAST](#)

[Megablast](#) compare short reads against htgs, nt, and wgs databases

[Parse blast XML output](#)

[Map with PerM for SOLiD and Illumina](#)

[Re-align with SRMA](#)

[Map with Mosaik](#)

NGS: Indel Analysis

NGS: RNA Analysis

NGS: SAM Tools

NGS: GATK Tools (beta)

NGS: Peak Calling

Bowtie2 (version 0.2)

Is this library mate-paired?:

Single-end ▾

FASTQ file:

4: FASTQ Groomer on data 3 ▾

Nucleotide-space: Must have Sanger-scaled quality values with ASCII

Write unaligned reads to separate file(s):

↙

Will you select a reference genome from your history or use a bu

Use one from the history ▾

Built-ins were indexed using default options

Select the reference genome:

14: <http://bioinfo.umassmed.edu/pub/data/filter/rRNA.fasta> ▾

Specify the read group for this file?:

No ▾

Parameter Settings:

Use defaults ▾

You can use the default

Execute

Tools

BAM

GS: SAM Tools

Convert from BAM to BED.

Find BAM alignments that overlap intervals in another file

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

flagstat provides simple stats on BAM files

Convert from BAM to BED. (version 0.1.0)

Convert the following BAM file to BED:

16: Bowtie2 on data 14 and data 4: aligned reads ▾

What type of BED output would you like:

Create a 6-column BED file. ▾

Report spliced BAM alignments as separate BED entries:

Use alignment's edit-distance for BED score:

Use other NUMERIC BAM alignment tag as the BED score:

Execute

for this file?:

Tools

BAM



GS: SAM Tools

Convert from BAM to BED.

Find BAM alignments that overlap intervals in another file

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

flagstat provides simple stats on BAM files

Convert from BAM to BED. (version 0.1.0)

Convert the following BAM file to BED:

16: Bowtie2 on data 14 and data 4: aligned reads



What type of BED output would you like:

Create a 6-column BED file.



Report spliced BAM alignments as separate BED entries:



Use alignment's edit-distance for BED score:



Use other NUMERIC BAM alignment tag as the BED score:

Execute

Group the Data and Count

Tools

Join, Subtract and Group

Join two Datasets side by side on a specified field

Compare two Datasets to find common or distinct rows

Subtract Whole Dataset from another dataset

Group data by a column and perform aggregate operation on other columns.

Column Join

Convert Formats

Extract Features

Fetch Sequences

Fetch Alignments

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Wavelet Analysis

Graph/Display Data

Regional Variation

Group (version 2.0.0)

Select data:

17: Bowtie2 on data 14 and data 4: aligned reads (as BED)

Dataset missing? See TIP below.

Group by column:

c1 ▲▼

Ignore case while grouping?:

Operations

Operation 1

Type:

Count ▲▼

On column:

c1 ▲▼

Round result to nearest integer?:

NO ▲▼

Rem

Analyze Data

Workflow

Shared Data ▾

Visualization ▾

Admin

Help ▾

User ▾

Add ne

3000072055985=5S#rRNA	10
gi 124517659 ref NR_003286.1 _Homo_sapiens_18S_ribosomal_RNA_(LOC100008588)	40
gi 124517661 ref NR_003287.1 _Homo_sapiens_28S_ribosomal_RNA_(LOC100008589)	1353
gi 142372596 ref NR_003285.2 _Homo_sapiens_5.8S_ribosomal_RNA_(LOC100008587)	32

Homework:

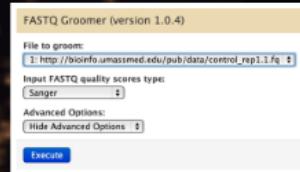
Galaxy tools for RNA-Seq Data

Read alignments and visualization

1. Upload



2. FastQ Groomer



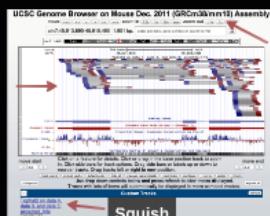
3. Map with Tophat2



4. Visualize BAM file



5. UCSC Genome Browser



6. Coverage Visualization

6a. Coverage



6b. bedGraph to bigWig conversion



6c. Visualization



6d. Browser



1. Upload

1a. Upload fastq files:

Upload File (version 1.1.3)

File Format:
fastq

Which format? See help below

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

URL/Text:
http://bioinfo.umassmed.edu/pub/data/control_rep1.1.fq
http://bioinfo.umassmed.edu/pub/data/control_rep1.2.fq

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
Your FTP upload directory contains no files.		

This Galaxy server allows you to upload files via FTP. To upload some files, log in to the Galaxy server at galaxy.umassmed.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:
Mouse Dec. 2011 (GRCm38/mm10) (mm10)

1b. Upload genome File:

<http://bioinfo.umassmed.edu/pub/data/mm10.fa> => fasta

2. FastQ Groomer

FASTQ Groomer (version 1.0.4)

File to groom:

1: http://bioinfo.umassmed.edu/pub/data/control_rep1.1.fq 

Input FASTQ quality scores type:

Sanger 

Advanced Options:

Hide Advanced Options 

Execute

3. Map with Tophat2

Tools

Evolution

Motif Tools

Multiple Alignments

Metagenomic analyses

FASTA manipulation

NGS: QC and manipulation

NGS: Mapping

NGS: Indel Analysis

NGS: RNA Analysis

RNA-SEQ

Tophat for Illumina Find splice junctions using RNA-seq data

Tophat2 Gapped-read mapper for RNA-seq data

Tophat for SOLiD 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

Cuffmerge merge together several

Tophat2 (version 0.5)

Is this library mate-paired?:

RNA-Seq FASTQ file, forward reads:

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

RNA-Seq FASTQ file, reverse reads:

Nucleotide-space: Must have Sanger-scaled quality values with ASCII offset 33

Mean Inner Distance between Mate Pairs:

Std. Dev for Distance between Mate Pairs:

The standard deviation for the distribution on inner distances between mate pairs.

Report discordant pair alignments?:

Use a built in reference genome or own from your history:

Built-ins genomes were created using default options

Select the reference genome:

TopHat settings to use:

History

70.7 MB

9: Tophat2 on data 4, data 3, and data 5: accepted_hits

8: Tophat2 on data 4, data 3, and data 5: splice junctions

7: Tophat2 on data 4, data 3, and data 5: deletions

6: Tophat2 on data 4, data 3, and data 5: insertions

5: http://bioinfo.umassmed.edu/pub/data/mm10.fa

4: FASTQ Groomer on data 2

3: FASTQ Groomer on data 1

2: http://bioinfo.umassmed.edu/pub/data/control_rep1.2.fq

1: http://bioinfo.umassmed.edu/pub/data/control_rep1.1.fq

4. Visualize BAM file

9: Tophat2 on data 4, data 3, and data 5: accepted hits

1.0 MB
format: bam, database: mm10
Log: tool progress Settings: Output
files: "genome.*.bt2" Line rate: 6 (line is 64 bytes) Lines per side: 1 (side is 64 bytes) Offset rate: 4 (one in 16)
FTable chars: 10 Strings: unpacked
Max bucket size: default Max bucket size, sqrt mu

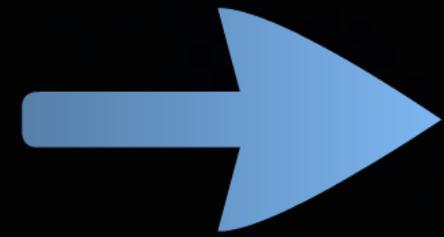
     

display at UCSC [main](#)
display in IGB [Local Web](#)

Binary bam alignments file

8: Tophat2 on data 4, data 3, and data 5: splice junctions

7: Tophat2 on data 4, data 3, and data 5: deletions



5. UCSC Genome Browser

UCSC Genome Browser on Mouse Dec. 2011 (GRCm38/mm10) Assembly

move <<< << < > >> zoom in 1.5x 3x 10x base zoom out 1.5x 3x go

chr7:45,613,890-45,615,490 1,601 bp. enter position, gene symbol or search terms

chr7 (qB3) 7qA1 A2 7qA3 7qB1 7qB3 7qB4 7qB5 7qC qD1 qD2 7qD3 7qE1 7qE3 7qF1 qF2 7qF3 F4qFS

scale
chr7:
500 bases | mm10
45,614,500 | 45,615,000
Tophat2 on data 4, data 3, and data 5: accepted_hits

Fgf21
Ensembl Genes 3,296
Ensembl Gene Predictions - Ensembl 71
Placental Cons 0
-3.94
Multiz Align Vertebrate Multiz Alignment & Conservation (60 Species)

move start < 2.0 > move end < 2.0 >
track search default tracks default order hide all manage custom tracks track hubs configure reverse resize refresh
collapse all expand all

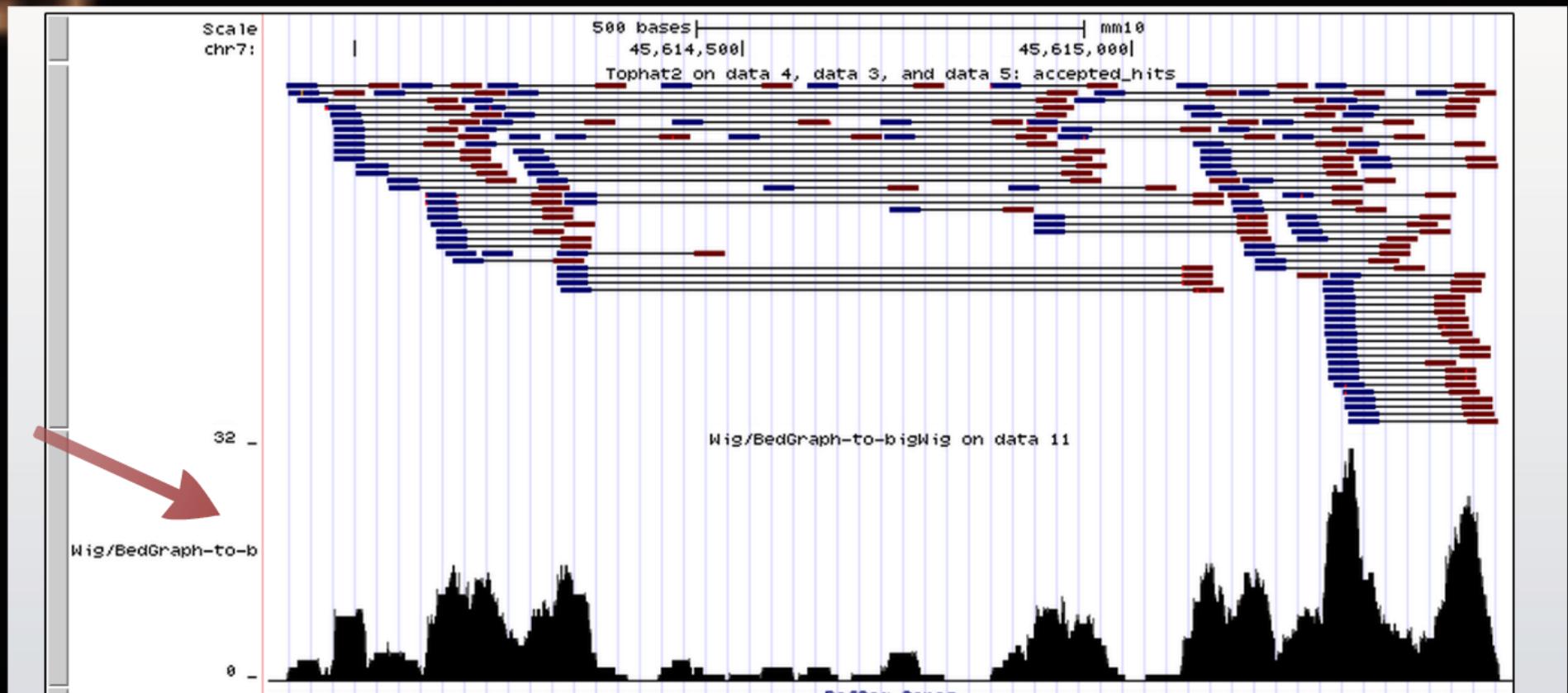
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

Custom Tracks refresh

-
Tophat2 on data 4,
data 3, and data 5:
accepted_hits
squish

Squish

6d. Browser



6. Coverage Visualization

6a. Coverage

Create a BedGraph of genome coverage (version 0.1.0)

The BAM or BED file from which coverage should be computed:
9: Tophat2 on data 4, data 3, and data 5: accepted_hits

Report regions with zero coverage:
 If set, regions without any coverage will also be reported.

Treat split/spliced BAM or BED12 entries as distinct BED intervals when computing coverage:
 If set, the coverage will be calculated based on the spliced intervals only. For BAM files, this uses the CIGAR N operation to infer the blocks for computing coverage. For BED12 files, this inspects the BlockCount, BlockStarts, and BlockEnds fields (i.e., columns 10, 11, 12). If this option is not set, the coverage will be calculated based on the interval's START/END coordinates, and would include introns of RNAseq data.

Calculate coverage based on:
 both strands combined

Scale the coverage by a constant factor:

Each BEDGRAPH coverage value is multiplied by this factor before being reported. Useful for scaling coverage by, e.g., reads per million (RPM).

Execute

6b. bedGraph to bigWig conversion

Wig/BedGraph-to-bigWig (version 1.1.0)

Convert:
11: Tophat2 on data 4, data 3, and data 5: accepted_hits (Genome Coverage BedGraph)

Converter settings to use:
 Default

Default settings should usually be used.

Execute

Syntax
This tool converts bedgraph or wiggle data into bigWig type.

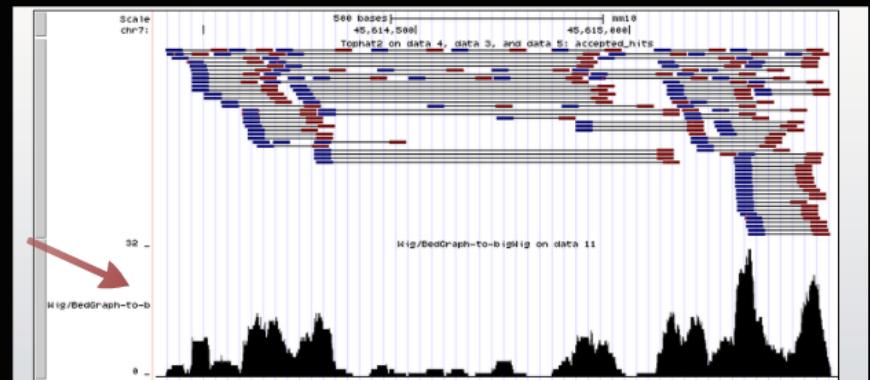
6c. Visualization

12: Wig/BedGraph-to-bigWig on data 11
227.4 KB
format: bigwig, database: mm10
 display at UCSC main
 display in IGB Local Web
Binary UCSC Bigwig file

11: Tophat2 on data 4, data 3, and data 5: accepted_hits (Genome Coverage BedGraph)
36,910 regions
format: bedgraph, database: mm10
 display at UCSC main

1. Chrom 2. Start 3. End 4
chr16 0 3006102 0
chr16 3006102 3006142 1
chr16 3006142 3006251 0
chr16 3006251 3006291 1
chr16 3006291 3011469 0

6d. Browser



6a. Coverage

Tools

Coverage 

Fetch Alignments

MAF Coverage Stats Alignment coverage information

Operate on Genomic Intervals

Base Coverage of all intervals

Coverage of a set of intervals on second set of intervals

Regional Variation

Feature coverage

NGS: SAM Tools

Create a histogram of genome coverage

Create a BedGraph of genome coverage 

Filter pileup on coverage and SNPs

NGS: GATK Tools (beta)

ALIGNMENT UTILITIES

Depth of Coverage on BAM files

Create a BedGraph of genome coverage (version 0.1.0)

The BAM or BED file from which coverage should be computed:
9: Tophat2 on data 4, data 3, and data 5: accepted_hits

Report regions with zero coverage:
If set, regions without any coverage will also be reported.

Treat split/spliced BAM or BED12 entries as distinct BED intervals when computing coverage:
If set, the coverage will be calculated based the spliced intervals only. For BAM files, this inspects the CIGAR N operation to infer the blocks for computing coverage. For BED12 files, this inspects the BlockCount, BlockStarts, and BlockEnds fields (i.e., columns 10,11,12). If this option is not set, coverage will be calculated based on the interval's START/END coordinates, and would include introns for RNAseq data.

Calculate coverage based on: both strands combined

Scale the coverage by a constant factor:
Each BEDGRAPH coverage value is multiplied by this factor before being reported. Useful for scaling coverage by, e.g., reads per million (RPM)

Execute

6b. bedGraph to bigWig conversion

Tools

bedGraph

Convert Formats

[GTF-to-BEDGraph converter](#)

[**Wig/BedGraph-to-bigWig converter**](#)

NGS: SAM Tools

[Merge multiple BedGraph files](#)

[Create a BedGraph of genome coverage](#)

Workflows

- [All workflows](#)

Wig/BedGraph-to-bigWig (version 1.1.0)

Convert:

11: Tophat2 on data 4, data 3, and data 5: accepted_hits (Genome Coverage Bed)

Converter settings to use:

Default settings should usually be used.

Syntax

This tool converts bedgraph or wiggle data into bigWig type.

6c. Visualization

12: Wig/BedGraph-to-bigWig on data 11

227.4 KB
format: bigwig, database: mm10

display at UCSC [main](#)

display in IGB [Local Web](#)

Binary UCSC BigWig file

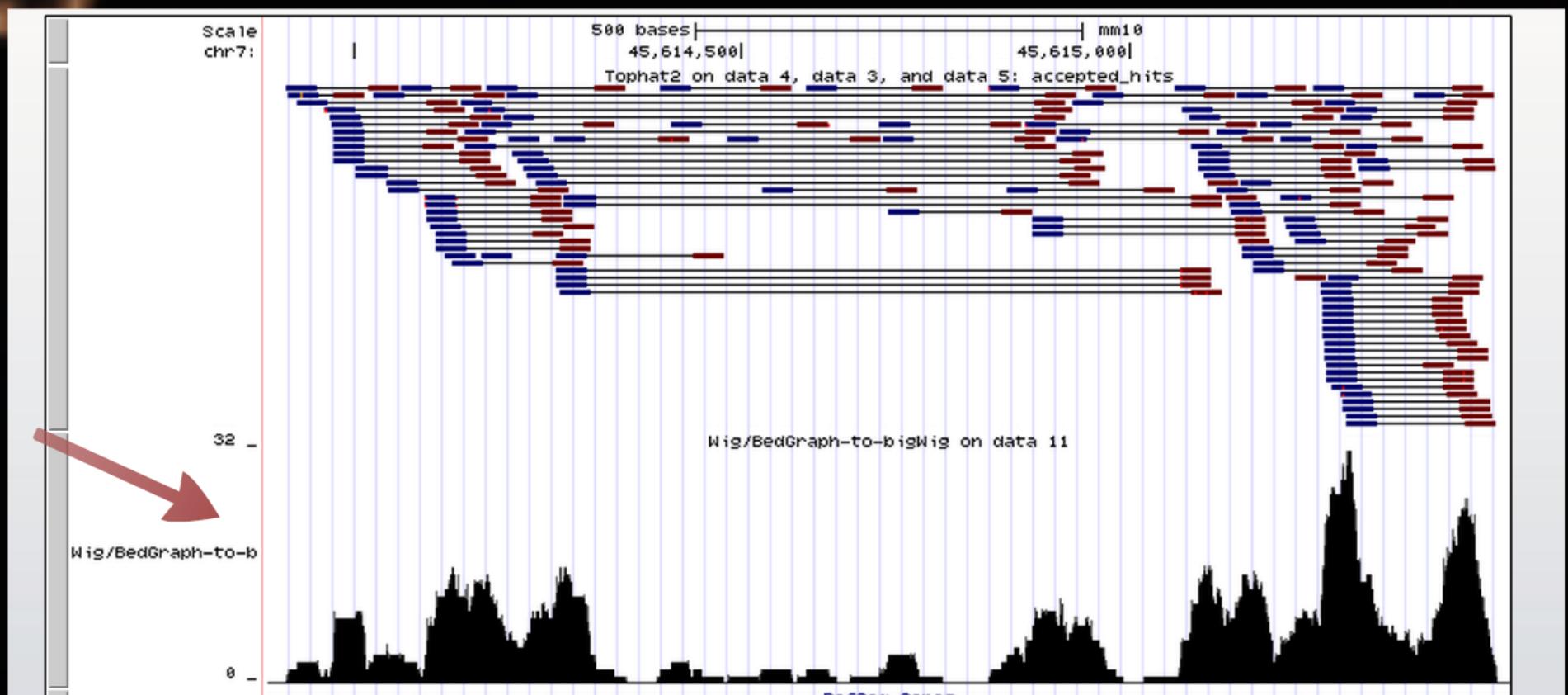
11: Tophat2 on data 4, data 3, and data 5: accepted_hits (Genome Coverage BedGraph)

36,910 regions
format: bedgraph, database: mm10

display at UCSC [main](#)

1.Chrom	2.Start	3.End	4
chr16	0	3006102	0
chr16	3006102	3006142	1
chr16	3006142	3006251	0
chr16	3006251	3006291	1
chr16	3006291	3011469	0

6d. Browser



Thanks!!!



© culture-nomad.com