

Galaxy Platform For NGS Data Analyses

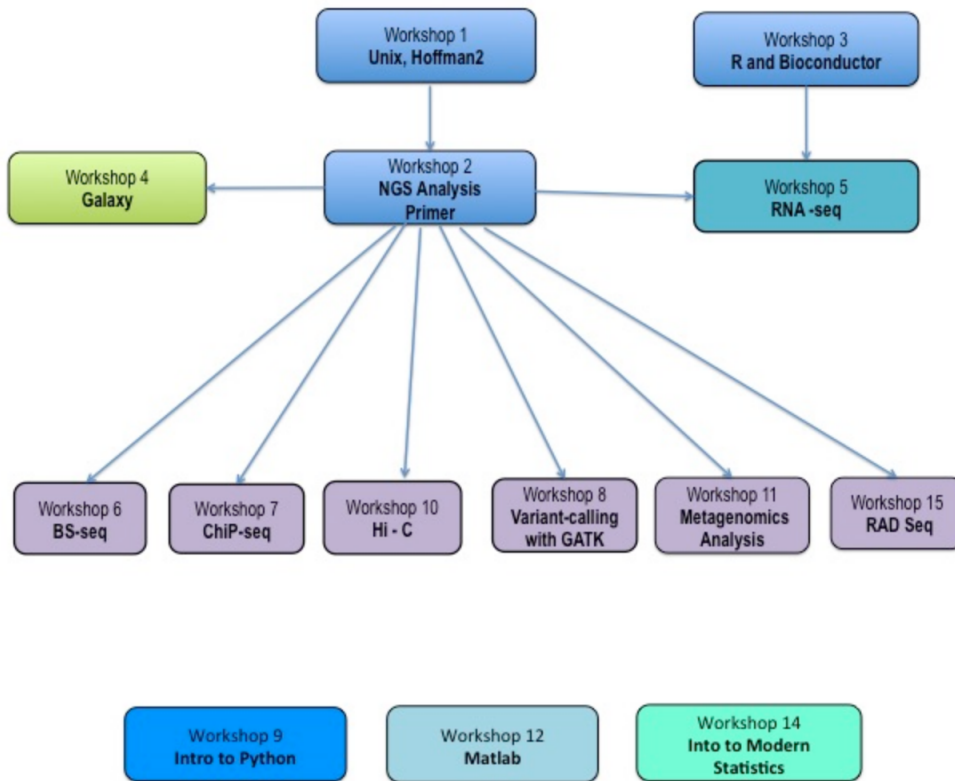
Weihong Yan

wyan@chem.ucla.edu

Collaboratory Web Site

<http://qcb.ucla.edu/collaboratory>

Collaboratory Workshops



W1: UNIX COMMAND LINE

W2: NGS ANALYSIS

W3: INTRO TO R

W4: GALAXY FOR NGS DATA
ANALYSIS

W5: RNA-SEQ ANALYSIS

W6: BS-SEQUENCING

W7: CHIP-SEQ ANALYSIS

W8: VARIANT DISCOVERY WITH GATK

W9: PYTHON

W 10: HI-C

W 11: METAGENOMICS ANALYSIS

W12: MATLAB

W14: INTRODUCTION TO MODERN
STATISTICS

W15: RAD-SEQ ANALYSIS

COLLABORATORY PYTHON USERS
GROUP

Workshop Outline

✓ Day 1

- UCLA galaxy and user account
- Galaxy web interface and management
- Tools for NGS analyses and their application
- Data formats
- Build/share workflow and history
- Q and A

✓ Day 2

- Galaxy Tools for RNA-seq analysis
- Galaxy Tools for ChIP-seq analysis
- Galaxy Tools for annotation.
- Q and A

*** Published datasets/results will be used in the tutorial

UCLA Galaxy

<http://galaxy.hoffman2.idre.ucla.edu>

✓ Hardware

- Headnode (1)
96Gb memory, 12 core
- Computing nodes (8)
48Gb memory, 12 core
- Storage
100 Tb disk space

✓ Galaxy Resource Management

- Hoffman2 grid engine
Default: 1 core/job
bowtie, bwa, tophat, cuffdiff, cufflinks, gatk programs: 4 core/job

UCLA Galaxy

<http://galaxy.hoffman2.idre.ucla.edu>

- ✓ galaxy login account:
login: your email associated with ucla
- ✓ Disk quota:
250Gb/user

Galaxy Account Management

The screenshot shows the Galaxy web interface in a browser window. The address bar displays `galaxy.hoffman2.idre.ucla.edu`. The top navigation bar includes links for Apple, iCloud, Facebook, Twitter, Wikipedia, Yahoo!, Popular, News, and font%3E. The main header shows "Galaxy / UCLA" and a navigation menu with "Analyze Data", "Workflow", "Shared Data", "Help", and "User". The "User" dropdown menu is open, showing options: "Logged in as mygalaxy@galaxy.ucla.edu", "Preferences", "Logout", "Saved Histories", "Saved Datasets", and "API Keys". The "User preferences" section is active, displaying the user's login status and a list of links: "Manage your information", "Change default permissions for new histories", "Manage your API keys", and "Logout of all user sessions". Below this, a message states: "You are using 13.5 Gb of disk space in this Galaxy instance. Your disk quota is expected? See the documentation for tips on how to find all of the data in your". On the left sidebar, a search bar is labeled "ch tools", and a list of tool categories is visible: "ita", "ver", "anipulation", "and Sort", "ubtract and Group", "rt Formats", "t Features", "Sequences", "nomic Scores", "te on Genomic Intervals", "ics", "onal Mutagenis", "ava genomics toolkit", "uantitation Tools", and "ols".

Galaxy

galaxy.hoffman2.idre.ucla.edu

Apple iCloud Facebook Twitter Wikipedia Yahoo! Popular News font%3E

Galaxy / UCLA

Analyze Data Workflow Shared Data Help User

User preferences

You are currently logged in as mygalaxy@galaxy.ucla.edu.

- [Manage your information](#)
- [Change default permissions](#) for new histories
- [Manage your API keys](#)
- [Logout](#) of all user sessions

You are using 13.5 Gb of disk space in this Galaxy instance. Your disk quota is expected? See the [documentation](#) for tips on how to find all of the data in your

Logged in as mygalaxy@galaxy.ucla.edu

Preferences

Logout

Saved Histories

Saved Datasets

API Keys

ch tools

ita

ver

anipulation

and Sort

ubtract and Group

rt Formats

t Features

Sequences

nomic Scores

te on Genomic Intervals

ics

onal Mutagenis

ava genomics toolkit

uantitation Tools

ols

The screenshot shows the Galaxy web interface at galaxy.hoffman2.idre.ucla.edu/root. The main content area displays the 'Upload File (version 1.1.3)' tool configuration. The left sidebar lists various tool categories under 'Tools', including 'Get Data', 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Insertional Mutagenis', 'NGS: java genomics toolkit', 'NGS: Quantitation Tools', 'GFF tools', 'EMBOSS', 'Motif Tools', 'FASTA manipulation', 'pyCRAC (work in progress)', 'Integrative Analysis', 'Human Genome Variation', 'Genome Diversity', 'NGS: QC and manipulation', 'NGS: Picard (beta)', 'NGS: Mapping', 'NGS: Methylation Mapping', 'NGS: Indel Analysis', 'NGS: RNA Analysis', 'NGS: BEDTools', 'NGS: SAM Tools', and 'NGS: GATK Tools (beta)'. The right sidebar shows the 'History' panel with a dropdown menu open, listing 'HISTORY LISTS' (Saved Histories, Histories Shared with Me), 'CURRENT HISTORY' (Create New, Clone, Copy Datasets, Share or Publish, Extract Workflow, Dataset Security, Show Deleted Datasets, Show Hidden Datasets, Purge Deleted Datasets, Show Structure, Export to File, Delete, Delete Permanently), and 'OTHER ACTIONS' (Import from File). The central tool configuration area includes fields for 'File Format' (Auto-detect), 'File' (Browse...), 'URL/Text' (text area), 'Files uploaded via FTP' (table with File, Size, Date columns), 'Convert spaces to tabs' (checkbox), 'Genome' (dropdown), and an 'Execute' button. The bottom of the interface has three blue arrows pointing to the tool list, the 'Execute' button, and the history panel respectively.

Installed tools

Launch analysis and view result

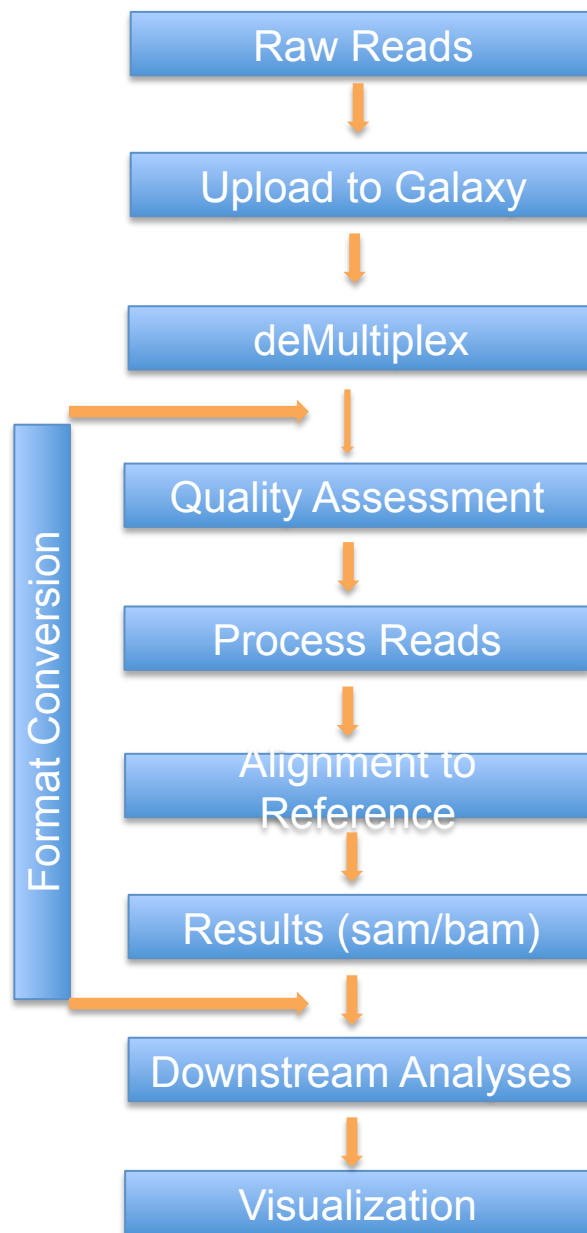
History of execution and results

Galaxy / UCLA

Tools ⚙

search tools

- [Get Data](#)
- [Lift-Over](#)
- [Text Manipulation](#)
- [Filter and Sort](#)
- [Join, Subtract and Group](#)
- [Convert Formats](#)
- [Extract Features](#)
- [Fetch Sequences](#)
- [Get Genomic Scores](#)
- [Operate on Genomic Intervals](#)
- [Statistics](#)
- [Insertional Mutagenis](#)
- [NGS: java genomics toolkit](#)
- [NGS: Quantitation Tools](#)
- [GFF tools](#)
- [EMBOSS](#)
- [Motif Tools](#)
- [FASTA manipulation](#)
- [Integrative Analysis](#)
- [Human Genome Variation](#)
- [Genome Diversity](#)
- [NGS: QC and manipulation](#)
- [NGS: Picard \(beta\)](#)
- [NGS: Mapping](#)
- [NGS: Methylation Mapping](#)
- [NGS: Indel Analysis](#)
- [NGS: RNA Analysis](#)
- [NGS: BEDTools](#)
- [NGS: SAM Tools](#)
- [NGS: GATK Tools \(beta\)](#)
- [NGS: Peak Calling](#)
- [NGS: Simulation](#)
- [Galaxy: Status](#)
- Workflows
 - [All workflows](#)



*_qseq.txt, *.fastq

File transfer protocol (ftp)

Barcode splitter, deMultiplex workflow

fastqc, compute quality statistics, draw quality score boxplot, draw nucleotides distribution

Trim sequences, sickle, scythe

bwa, bowtie, bowtie2, tophat

Text manipulation toolkit, BEDTools, SAM Tools, java genomics toolkit, picard toolkit

BS-Seeker2, cufflinks, cuffdiff, macs, macs2, GATK, CEAS

Genome browser, IGV

Repositories of Galaxy Tools

<https://toolshed.g2.bx.psu.edu>

Galaxy Tool Shed

RepositoriesHelpUser

2957 valid tools on Jan 30, 2015

Search

- [Search for valid tools](#)
- [Search for workflows](#)

Valid Galaxy Utilities

- [Tools](#)
- [Custom datatypes](#)
- [Repository dependency definitions](#)
- [Tool dependency definitions](#)

All Repositories

- [Browse by category](#)

Available Actions

- [Login to create a repository](#)

Repositories by Category

search repository name, description

Name	Description
Assembly	Tools for working with assemblies
ChIP-seq	Tools for analyzing and manipulating ChIP-seq data.
Combinatorial Selections	Tools for combinatorial selection
Computational chemistry	Tools for use in computational chemistry
Convert Formats	Tools for converting data formats
Data Managers	Utilities for Managing Galaxy's built-in data cache
Data Source	Tools for retrieving data from external data sources
Fasta Manipulation	Tools for manipulating fasta data
Fastq Manipulation	Tools for manipulating fastq data
Genome-Wide Association Study	Utilities to support Genome-wide association studies
Genomic Interval Operations	Tools for operating on genomic intervals
Graphics	Tools producing images
Imaging	Utilities to support imaging
Metabolomics	Tools for use in the study of Metabolomics
Metagenomics	Tools enabling the study of metagenomes
Micro-array Analysis	Tools for performing micro-array analysis
Next Gen Mappers	Tools for the analysis and handling of Next Gen sequencing data
Ontology Manipulation	Tools for manipulating ontologies
Phylogenetics	Tools for performing phylogenetic analysis
Proteomics	Tools enabling the study of proteins
RNA	Utilities for RNA
SAM	Tools for manipulating alignments in the SAM format

Using 1%

History

day1

2: UCSC Main
knownGene
~1,500,000
format: gtf
display at

1. Seqname
chr1
chr1
chr1
chr1
chr1
chr1

1: UCSC Main
knownGene
82,960 regions
format: bed, database: hg19
display at UCSC [main](#)
view in [GeneTrack](#)
display at RViewer [main](#)

1. Chrom	2. Start	3. End	4. Name	5	6
chr1	11873	14409	uc001aaa.3	0	+
chr1	11873	14409	uc010nxr.1	0	+
chr1	11873	14409	uc010nxq.1	0	+
chr1	14361	16765	uc009vis.3	0	-
chr1	16857	17751	uc009vjc.1	0	-
chr1	15795	18061	uc009vjd.2	0	-

HISTORY LISTS

- Saved Histories
- Histories Shared with Me

CURRENT HISTORY

- Create New
- Clone
- Copy Datasets
- Share or Publish
- Extract Workflow
- Dataset Security
- Show Deleted Datasets
- Show Hidden Datasets
- Purge Deleted Datasets
- Show Structure
- Export to File
- Delete
- Delete Permanently

OTHER ACTIONS

- Import from File

Edit Attributes

Name:
knownGene

Info:
Download from UCSC browser

Annotation / Notes:
release date

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

Database/Build:
Human Feb. 2009 (GRCh37/hg19) (hg19)

Number of comment lines:
☐

Save

Auto-detect
This will inspect the dataset and attempt to correct the above column values if they are not accurate.

Convert to new format

Convert GFF to BED

This will create a new dataset with the contents of this dataset converted to a new format.

Convert

Change data type

New Type:
gtf

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

Save

Manage dataset permissions on UCSC Main on Human: knownGene (genome)

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

Roles associated:
mygalaxy@galaxy.ucla.edu

- ✓ History panel contains all datasets that are uploaded and results derived from certain analyses
- ✓ A history can be organized, annotated, and managed as a project
- ✓ History is sharable.
- ✓ Workflow is extracted and built from a history
- ✓ Each dataset under a history can be viewed, examined, converted to other formats, and annotated.

Getting Data to the Galaxy

The screenshot shows the Galaxy web interface at galaxy.hoffman2.idre.ucla.edu. The main panel displays the 'Upload File (version 1.1.3)' tool. On the left, the 'Tools' sidebar lists various categories, with 'UCSC Main table browser' and 'UCSC Archaea table browser' highlighted by blue arrows. The 'Upload File' tool form includes a 'File Format' dropdown set to 'Auto-detect', a 'File' section with a 'Browse...' button and a text area containing a URL, and a 'URL/Text' section with a text area containing a URL. A blue arrow points to the 'Browse...' button, and another points to the URL text area. Below the 'URL/Text' section, a table titled 'Files uploaded via FTP' shows no files. A blue arrow points to this table. The 'Convert spaces to tabs' section has a 'Yes' radio button selected. The 'Genome' dropdown is set to 'unspecified (?)'. An 'Execute' button is at the bottom. On the right, the 'History' panel shows 'Unnamed history' with 0 bytes and a message: 'Your history is empty. Click 'Get Data' on the left pane to start'.

Galaxy / UCLA

Analyze Data Workflow Shared Data Help User

Using 0%

Tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Insertional Mutagenesis

NGS: java genomics toolkit

NGS: Quantitation Tools

GFF tools

EMBOSS

Motif Tools

FASTA manipulation

pyCRAC (work in progress)

Integrative Analysis

Human Genome Variation

Genome Diversity

NGS: QC and manipulation

NGS: Picard (beta)

NGS: Mapping

Upload File (version 1.1.3)

File Format: Auto-detect

Which format? See help below

File: Browse... No file selected.

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:

http://hgdownload.soe.ucsc.edu/goldenPath/danRer7/bigZips/mrna.fa.gz

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 FTP server at galaxy.hoffman2.idre.ucla.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: unspecified (?)

Execute

History

Unnamed history 0 bytes

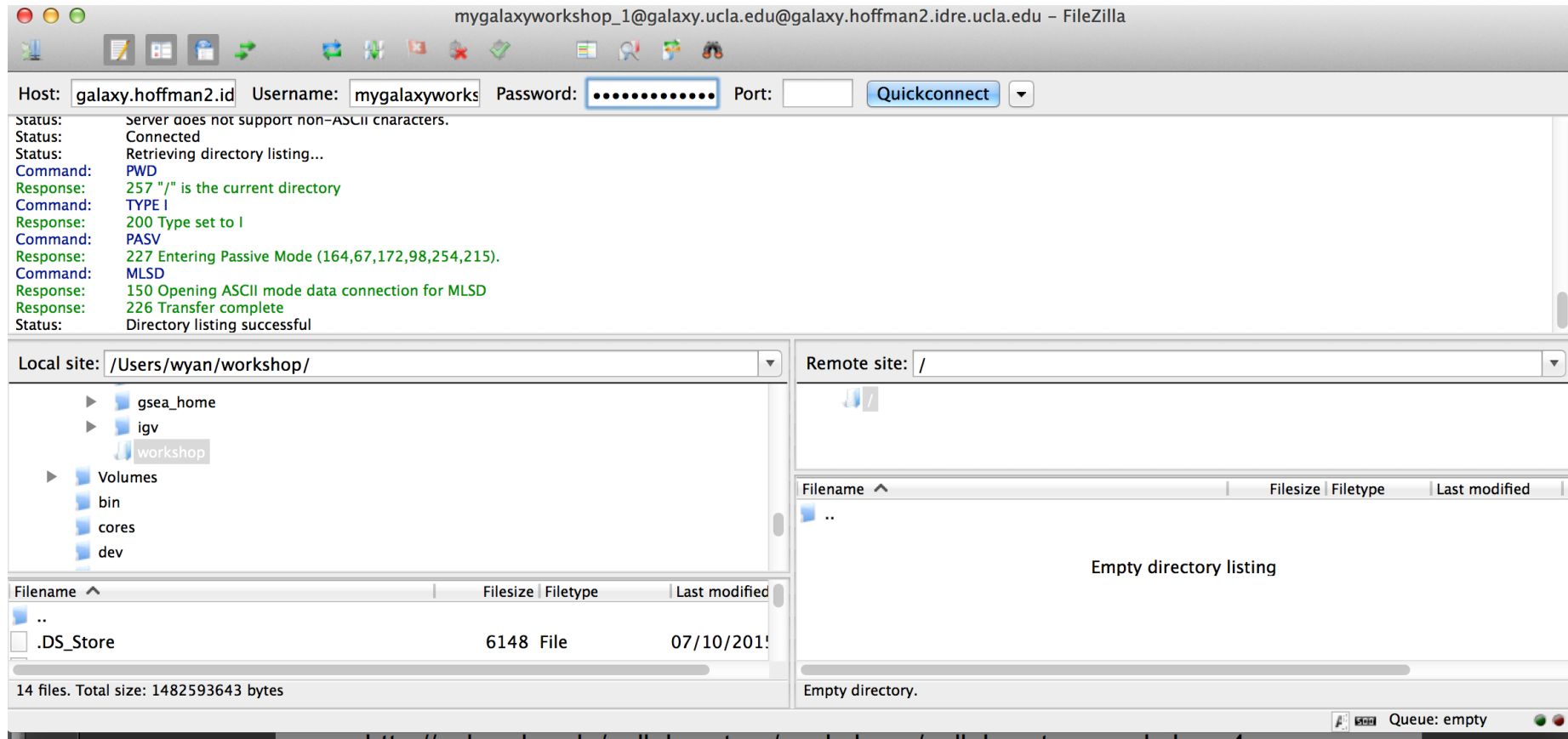
Your history is empty. Click 'Get Data' on the left pane to start

UCSC table browser: allows to upload genome assembly and annotations to the galaxy

Data libraries: datasets need to be put on the galaxy server before they can be uploaded.

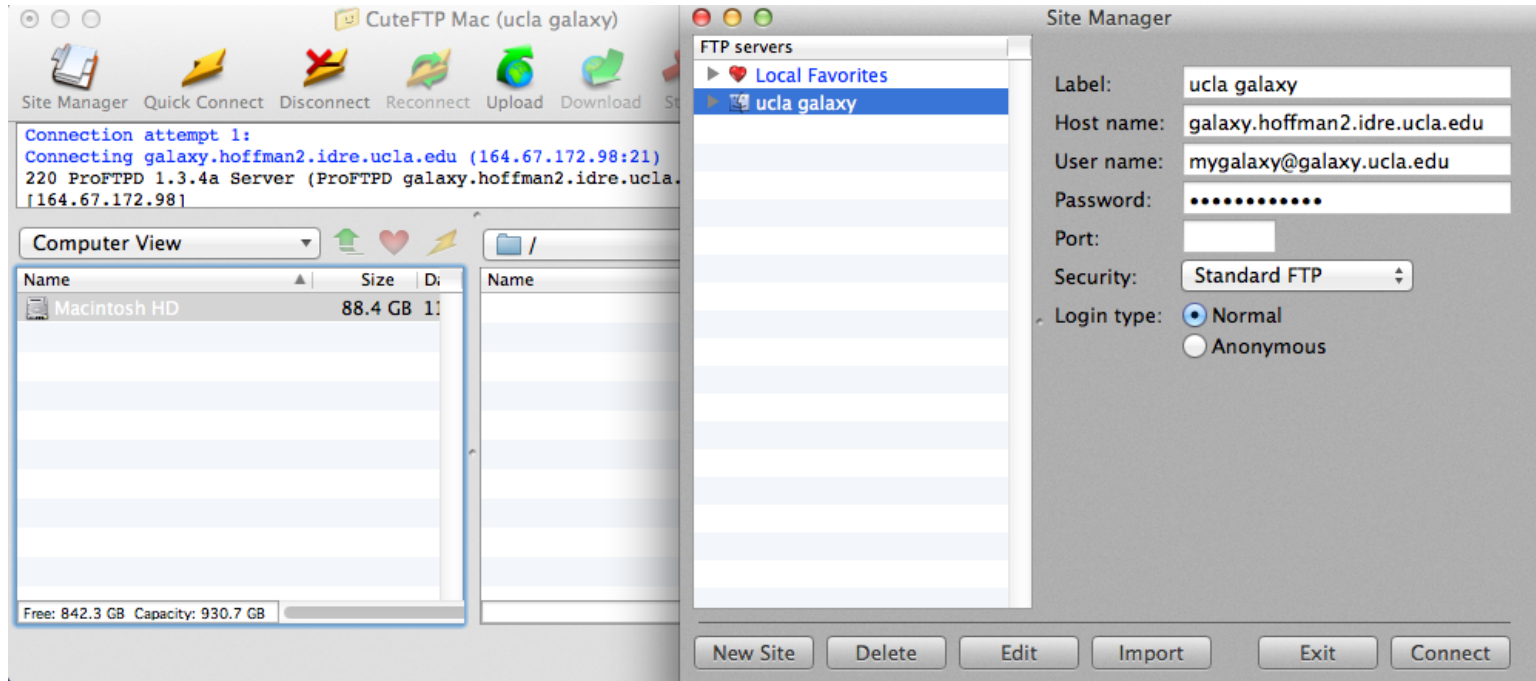
(Secure) FTP Clients

FileZilla: <http://filezilla-project.org>



Host: galaxy.hoffman2.idre.ucla.edu
Username and Password: galaxy login

(Secure) FTP Clients



<https://www.bol.ucla.edu/software/mac/cuteftp/>

Upload Data to Galaxy

✓ Data Transfer from your Hoffman2 Account"

```
[wyman@login3 ~]$ module load gftp
[wyman@login3 ~]$ gftp galaxy.hoffman2.idre.ucla.edu
gFTP 2.0.19, Copyright (C) 1998-2008 Brian Masney <masneyb@gftp.org>. If you have any questions, comments, or suggestions about this
program, please feel free to email them to me. You can always find out the latest news about gFTP from my website at
http://www.gftp.org/
gFTP comes with ABSOLUTELY NO WARRANTY; for details, see the COPYING file. This is free software, and you are welcome to
redistribute it under certain conditions; for details, see the COPYING file
```

```
Username [anonymous]: mygalaxy@galaxy.ucla.edu
```

```
Password:
```

```
Looking up galaxy.hoffman2.idre.ucla.edu
```

```
Trying galaxy.hoffman2.idre.ucla.edu:21
```

```
Connected to galaxy.hoffman2.idre.ucla.edu:21
```

```
220 ProFTPD 1.3.4a Server (ProFTPD galaxy.hoffman2.idre.ucla.edu FTP) [164.67.172.98]
```

```
USER mygalaxy@galaxy.ucla.edu
```

```
331 Password required for mygalaxy@galaxy.ucla.edu
```

```
PASS xxxx
```

```
230 User mygalaxy@galaxy.ucla.edu logged in
```

```
SYST
```

```
215 UNIX Type: L8
```

```
TYPE A
```

```
200 Type set to A
```

```
PWD
```

```
257 "/" is the current directory
```

```
ftp> █
```

[wyan@login1 ~]\$ gftp galaxy.hoffman2.idre.ucla.edu
gFTP 2.0.19, Copyright (C) 1998-2008 Brian Masney <masneyb@gftp.org>. If you have any questions, comments, or suggestions about this program, please feel free to email them to me. You can always find out the latest news about gFTP from my website at <http://www.gftp.org/>
gFTP comes with ABSOLUTELY NO WARRANTY; for details, see the COPYING file. This is free software, and you are welcome to redistribute it under certain conditions; for details, see the COPYING file

Username [anonymous]: wyan@chem.ucla.edu

Password:

Looking up galaxy.hoffman2.idre.ucla.edu

Trying galaxy.hoffman2.idre.ucla.edu:21

Connected to galaxy.hoffman2.idre.ucla.edu:21

220 ProFTPD 1.3.4a Server (ProFTPD galaxy.hoffman2.idre.ucla.edu FTP) [164.67.172.98]

USER wyan@chem.ucla.edu

331 Password required for wyan@chem.ucla.edu

PASS xxxx

230 User wyan@chem.ucla.edu logged in

SYST

215 UNIX Type: L8

TYPE A

200 Type set to A

PWD

257 "/" is the current directory

ftp> help

Supported commands:

about	clear	lcd	lmkdir	mkdir	rename
ascii	close	lchdir	lpwd	mput	rmdir
binary	delete	lchmod	lrename	open	set
cd	dir	ldelete	lrmdir	put	site
chdir	get	ldir	ls	pwd	
chmod	help	lls	mget	quit	

ftp> help put

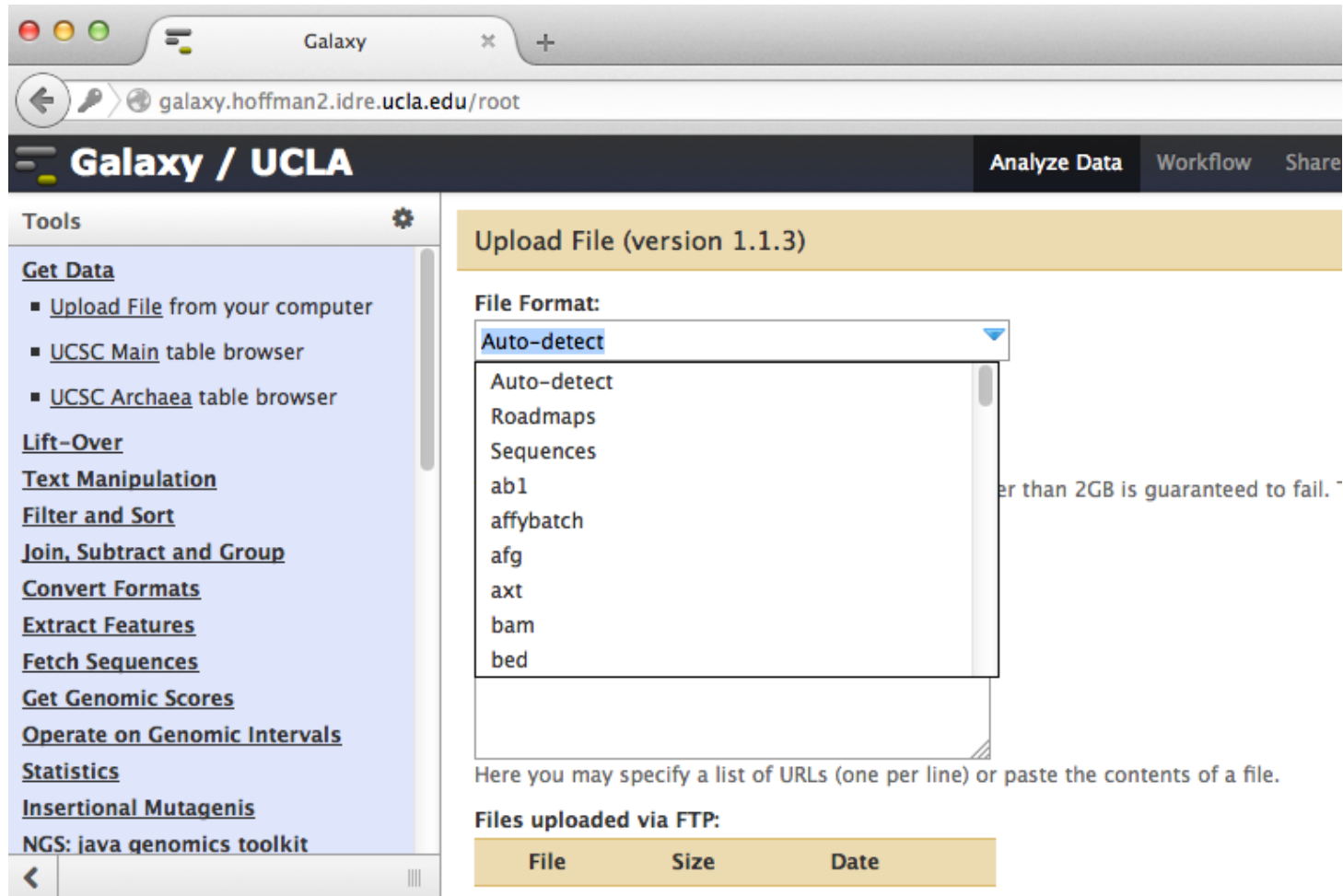
Uploads local file(s)

ftp> help quit

Exit from gFTP

ftp> put filename

File Formats



The screenshot shows the Galaxy web interface in a browser window. The address bar displays `galaxy.hoffman2.idre.ucla.edu/root`. The main header includes the 'Galaxy / UCLA' logo and navigation links for 'Analyze Data', 'Workflow', and 'Share'. On the left, a 'Tools' sidebar lists various categories: 'Get Data' (with sub-items like 'Upload File from your computer'), 'Lift-Over', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Convert Formats', 'Extract Features', 'Fetch Sequences', 'Get Genomic Scores', 'Operate on Genomic Intervals', 'Statistics', 'Insertional Mutagenis', and 'NGS: java genomics toolkit'. The main content area is titled 'Upload File (version 1.1.3)'. It features a 'File Format:' dropdown menu currently set to 'Auto-detect', with a list of other formats including Roadmaps, Sequences, ab1, affybatch, afg, axt, bam, and bed. Below the dropdown, there is a text input field with the instruction: 'Here you may specify a list of URLs (one per line) or paste the contents of a file.' At the bottom, a section titled 'Files uploaded via FTP:' contains a table with columns for 'File', 'Size', and 'Date'.

Galaxy / UCLA

Analyze Data Workflow Share

Tools

Get Data

- Upload File from your computer
- UCSC Main table browser
- UCSC Archaea table browser

Lift-Over

Text Manipulation

Filter and Sort

Join, Subtract and Group

Convert Formats

Extract Features

Fetch Sequences

Get Genomic Scores

Operate on Genomic Intervals

Statistics

Insertional Mutagenis

NGS: java genomics toolkit

Upload File (version 1.1.3)

File Format:

Auto-detect

- Auto-detect
- Roadmaps
- Sequences
- ab1
- affybatch
- afg
- axt
- bam
- bed

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

Files uploaded via FTP:

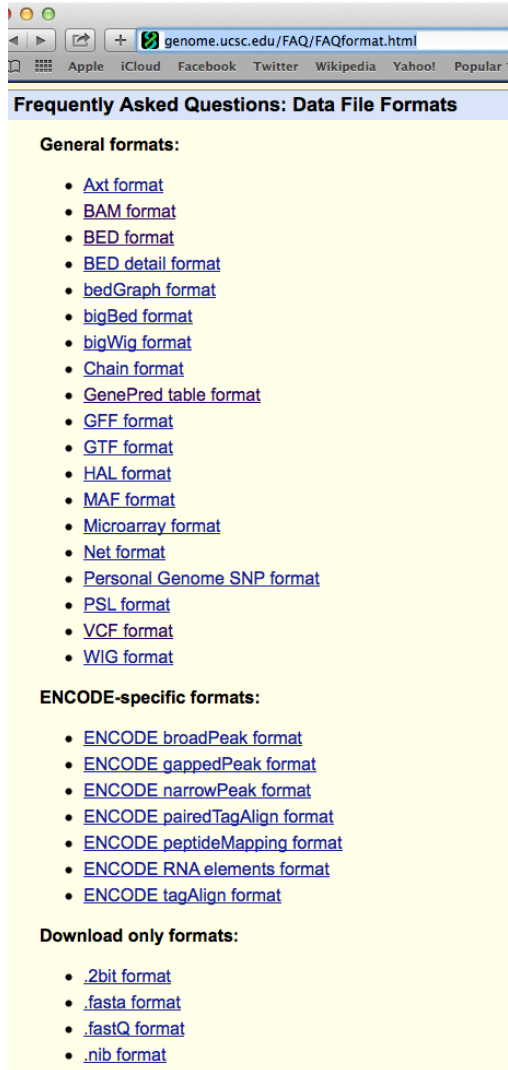
File	Size	Date
------	------	------

File Formats

- ✓ Formats created by application
 - roadmaps from assembler Velvet, gatk_dbsnp, gatk_recal...from GATK, lav and axt from blastz...
- ✓ Formats used for sequences and sequencing qualities
 - fasta, fastq, fastqSolexa, fastqillumina, fastqsanger...
- ✓ Formats used for annotations
 - BED (bigBed), GFF (general feature format), GFF3, GTF (gene transfer format), GenePred
- ✓ Formats used for NGS alignment information
 - sam (sequence alignment/map), bam (compressed binary version of sam)
- ✓ Formats used for displaying continuous-valued data
 - wig (wiggle), bigWig (indexed binary format of WIG), bedGraph
- ✓ Formats for variation data
 - vcf (variant call format), pgSnp (personal genome SNP format)

File Formats

<http://genome.ucsc.edu/FAQ/FAQformat.html>



The screenshot shows a web browser window with the address bar displaying genome.ucsc.edu/FAQ/FAQformat.html. The page title is "Frequently Asked Questions: Data File Formats". Under the "General formats:" section, there is a list of 20 file formats with links. Under the "ENCODE-specific formats:" section, there is a list of 8 file formats with links. Under the "Download only formats:" section, there is a list of 4 file formats with links.

Frequently Asked Questions: Data File Formats

General formats:

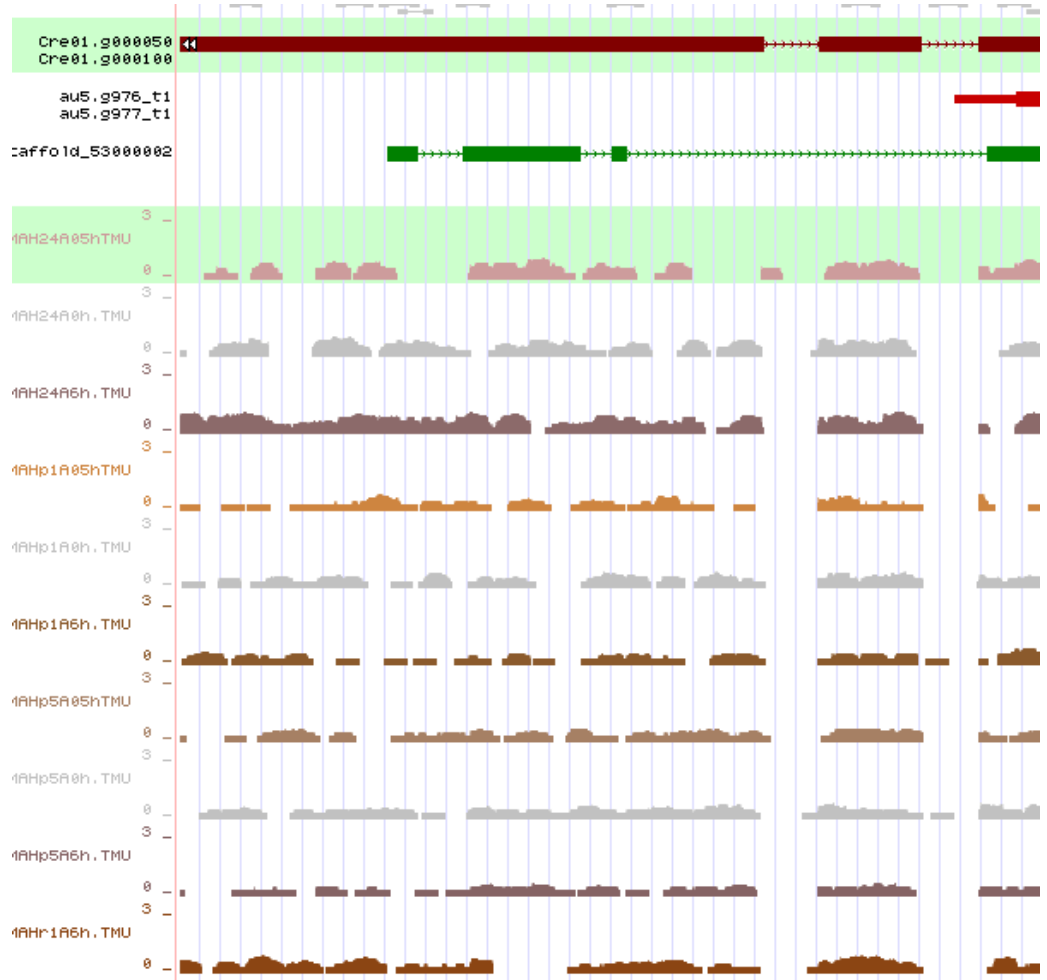
- [Axt format](#)
- [BAM format](#)
- [BED format](#)
- [BED detail format](#)
- [bedGraph format](#)
- [bigBed format](#)
- [bigWig format](#)
- [Chain format](#)
- [GenePred table format](#)
- [GFF format](#)
- [GTF format](#)
- [HAL format](#)
- [MAF format](#)
- [Microarray format](#)
- [Net format](#)
- [Personal Genome SNP format](#)
- [PSL format](#)
- [VCF format](#)
- [WIG format](#)

ENCODE-specific formats:

- [ENCODE broadPeak format](#)
- [ENCODE gappedPeak format](#)
- [ENCODE narrowPeak format](#)
- [ENCODE pairedTagAlign format](#)
- [ENCODE peptideMapping format](#)
- [ENCODE RNA elements format](#)
- [ENCODE tagAlign format](#)

Download only formats:

- [.2bit format](#)
- [.fasta format](#)
- [.fastQ format](#)
- [.nib format](#)



Retrieving Data from UCSC

The screenshot displays the Galaxy web interface at galaxy.hoffman2.idre.ucla.edu/. The main tool being used is the **Table Browser**, which is configured to retrieve data from the UCSC Genomes track. The configuration includes:

- clade:** Mammal
- genome:** Human
- assembly:** Feb. 2009 (GRCh37/hg19)
- group:** Genes and Gene Predictions
- track:** UCSC Genes
- table:** knownGene
- region:** genome (selected), chr1:981104-1068301
- identifiers (names/accessions):** paste 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

Buttons for **get output** and **summary/statistics** are visible. A link to [Using the Table Browser](#) is provided at the bottom of the configuration panel.

On the right side, the **History** panel shows two recent jobs:

- 2: UCSC Main on Human: knownGene (genome)**
~1,500,000 lines
format: gtf, database: hg19
display at UCSC [main](#)
- 1: UCSC Main on Human: knownGene (genome)**
82,960 regions
format: bed, database: hg19
display at UCSC [main](#)
view in [GeneTrack](#)
display at RViewer [main](#)

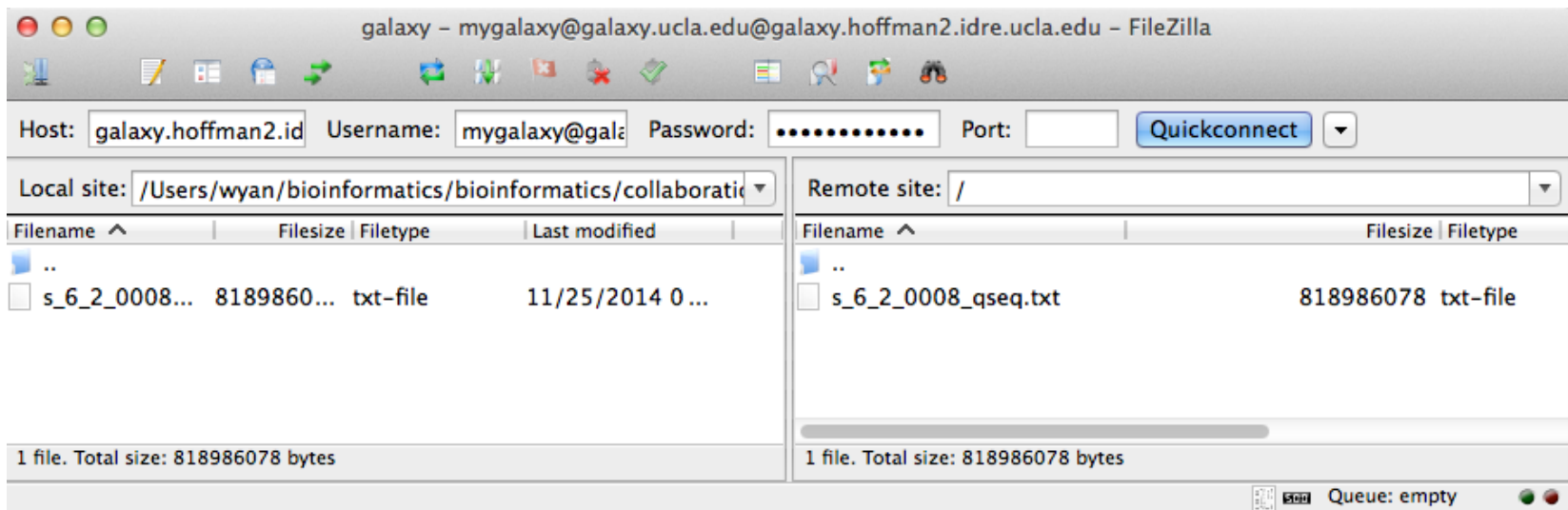
The bottom table in the History panel shows a list of genomic features:

1. Chrom	2. Start	3. End	4. Name	5
chr1	11873	14409	uc001aaa.3	0
chr1	11873	14409	uc010nxx.1	0
chr1	11873	14409	uc010nxq.1	0
chr1	14361	16765	uc009vis.3	0
chr1	16857	17751	uc009vjc.1	0
chr1	15795	18061	uc009vjd.2	0

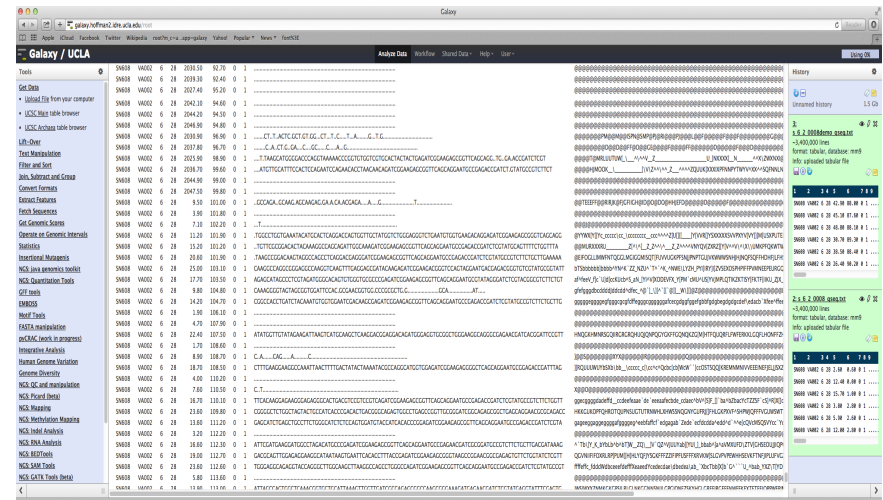
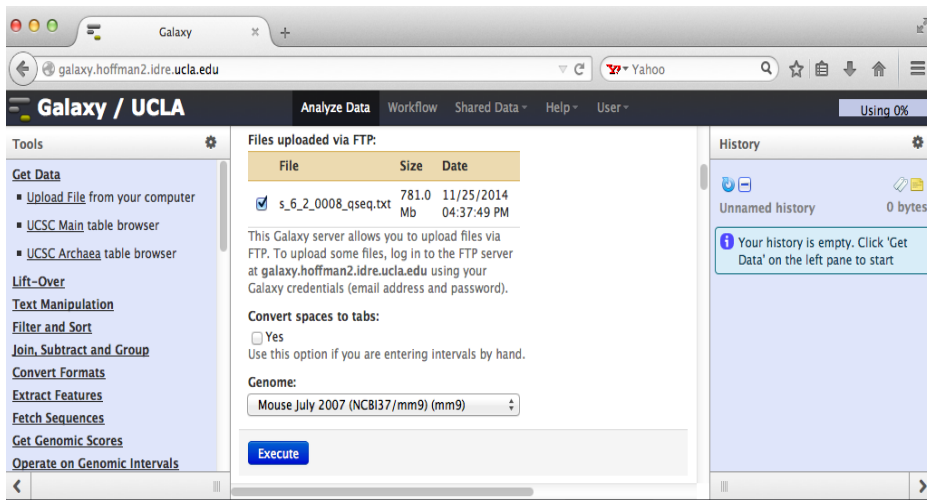
Retrieve knownGene table in two formats from UCSC genome site

Genomes Pre-installed in Galaxy

A. thaliana Jan. 2009 (TAIR9) (araTha2)
C. elegans Aug. 2007 (WS180/ce5) (ce5)
D. melanogaster Apr. 2006 (BDGP R5/dm3) (dm3)
Dog Sep 2011 (Broad/canFam3) (canFam3)
Horse Sep. 2007 (Broad/equCab2) (equCab2)
Human Feb. 2009 (GRCh37/hg19) (hg19)
Human Mar. 2006 (NCBI36/hg18) (hg18)
M. fascicularis Jul. 2011 (gigadb/CE) (macFas)
Mouse Dec. 2011 (GRCm38/mm10) (mm10)
Mouse July 2007 (NCBI37/mm9) (mm9)
R. norvegicus Mar. 2012 (RGSC 5.0/rn5) (rn5)
S. cerevisiae Apr. 2011 (SGD/sacCer3) (sacCer3)
S. cerevisiae June 2008 (SGD/sacCer2) (sacCer2)
X. laevis Version 7.1 (xenbase 3.1.1/Xenla_7.1_JGI) (xenLae1)
X. tropicalis Nov. 2009 (JGI 4.2/xenTro3) (xenTro3)



upload s_1_1_600000_qseq.txt to galaxy or use published qc history



qseq file format

- ✓ a plain-text file format for sequence reads.
- ✓ Each line contains: sequencer identifier, run number, lane number, tile number, x coordinate, y coordinate, index, read number (1 for single, 1 or 2 for paired ends), sequence, quality, filter

FastQ File Format

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

A FASTQ file normally uses four lines per sequence.

- Line 1 begins with a '@' character and is followed by a sequence identifier and an *optional* description (like a [FASTA](#) title line).
- Line 2 is the raw sequence letters.
- Line 3 begins with a '+' character and is *optionally* followed by the same sequence identifier (and any description) again.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

A FASTQ file containing a single sequence might look like this:

```
@SEQ_ID
GATTGGGGTTCAAAGCAGTATCGATCAATAGTAAATCCATTGTTCAACTCACAGTTT
+
!''*(((((***+))%%&&+))%%&&).1***-+*'')**55CCF>>>>>CCCCCCC65
```

The character '!' represents the lowest quality while '~' is the highest. Here are the quality value characters in left-to-right increasing order of quality ([ASCII](#)):

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNPOQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

FASTQ files from the [NCBI/EBI Sequence Read Archive](#) often include a description, e.g.

```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

FastQ Quality Scores



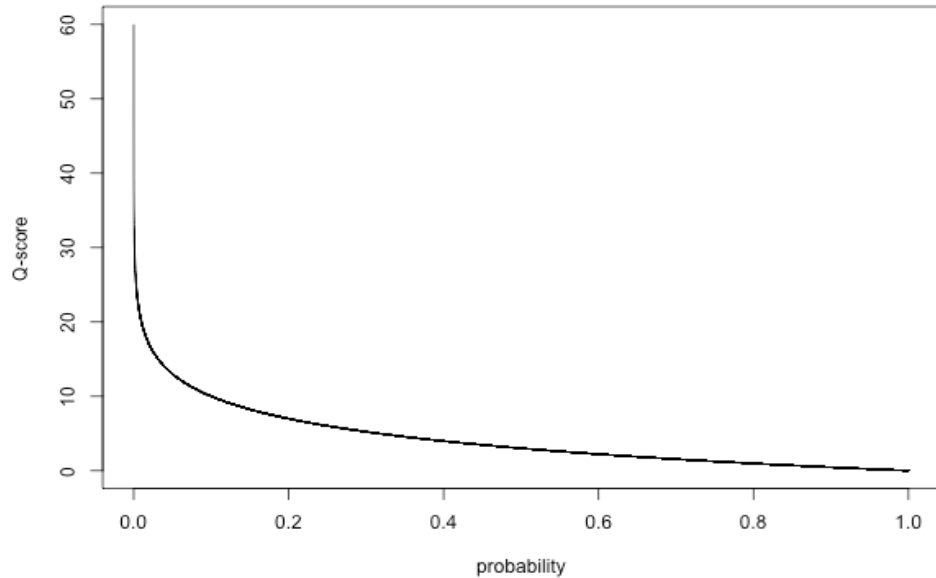
```

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)

```

Phred Quality Score

$$Q_{\text{sanger}} = -10 \log_{10} p$$



Phred quality score	P. That the base is called wrong	Accuracy of the base call
10	1 in 10	90%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Expected Sequence Quality

- ✓ A good quality read will have quality scores all above 28.

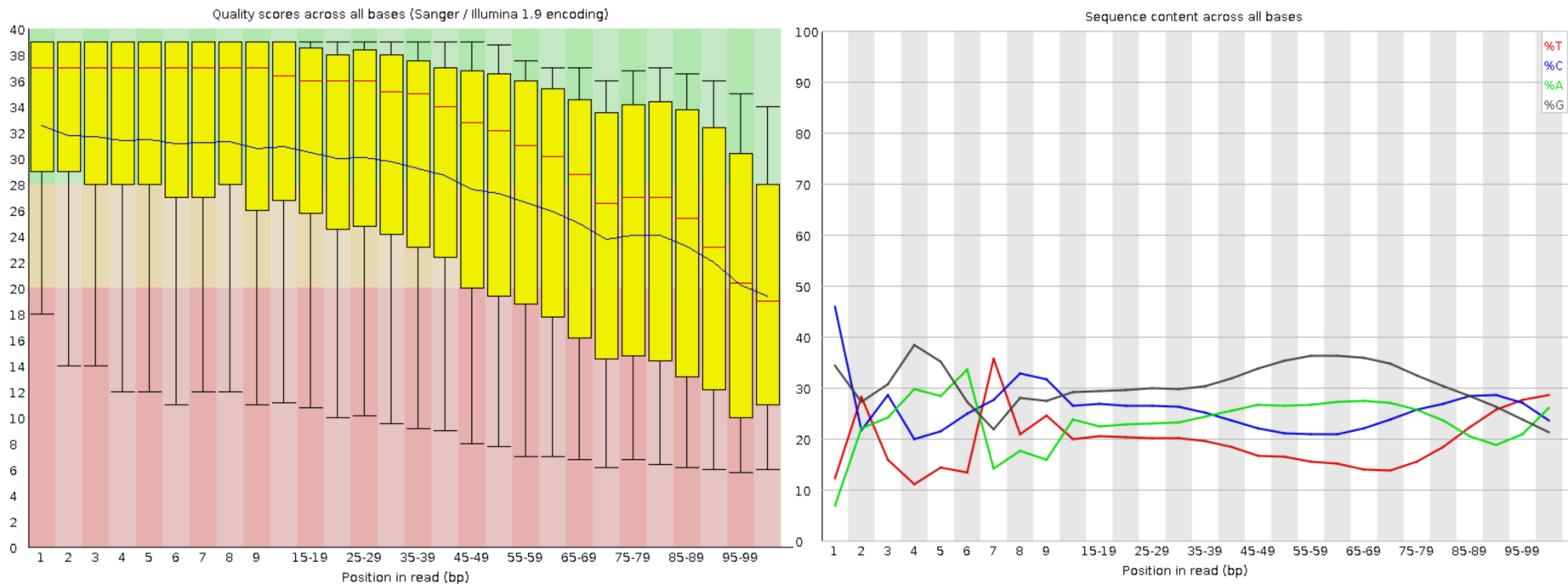
Trim reads with lower quality score.

- ✓ Per base sequence and GC content

Ideal reads have no variation with GC content along the length of the read.

Quality Control of Raw Sequences

- ✓ Upload s_1_1_600000_qseq.txt
- ✓ Run qseq_to_fastq program
- ✓ Run Fastqc program



Alternatively, use compute quality statistics -> draw quality score boxplot -> draw nucleotides distribution chart programs

FastQ Converter

[illegible]

FastQ Manipulation

The screenshot shows the Galaxy web interface with the 'Sickle (version 1.0.0)' tool selected. The left sidebar contains a list of tools under 'Tools' and 'Quality format converter (ASCII-ROCHE-454 DATA)'. The main panel displays the tool's configuration options: 'Single-End or Paired-End reads?' (set to 'Single-End'), 'Single-End FastQ Reads' (set to '2: FASTQ file'), 'Quality Threshold' (set to '20'), 'Length Threshold' (set to '20'), 'Don't do 5' trimming' (unchecked), and 'Discard sequences with Ns' (unchecked). An 'Execute' button is visible. Below the configuration, a detailed description of the Sickle tool is provided, explaining its sliding window algorithm and quality thresholds.

Tools

- 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
- 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
- Sickle Windowed Adaptive Trimming of FastQ data
- Scythe Trimming adapters/contaminants using a Naive Bayesian classifier
- AB-SOLID DATA
- Quality format converter (ASCII-

Sickle (version 1.0.0)

Single-End or Paired-End reads?:
Single-End

Note: Sickle will infer the quality type of the file from its datatype. I.e., if the datatype is fastqsanger, then the quality type is sanger. The default is fastqsanger.

Single-End FastQ Reads:
2: FASTQ file

Quality Threshold:
20

Length Threshold:
20

Don't do 5' trimming:
☐

Discard sequences with Ns:
☐

Execute

Most modern sequencing technologies produce reads that have deteriorating quality towards the 3'-end and some towards the 5'-end as well. Incorrectly called bases in both regions negatively impact assemblies, mapping, and downstream bioinformatics analyses.

Sickle is a tool that uses sliding windows along with quality and length thresholds to determine when quality is sufficiently low to trim the 3'-end of reads and also determines when the quality is sufficiently high enough to trim the 5'-end of reads. It will also discard reads based upon the length threshold. It takes the quality values and slides a window across them whose length is 0.1 times the length of the read. If this length is less than 1, then the window is set to be equal to the length of the read. Otherwise, the window slides along the quality values until the average quality in the window rises above the threshold, at which point the algorithm determines where within the window the rise occurs and cuts the read and quality there for the 5'-end cut. Then when the average quality in the window drops below the threshold, the algorithm determines where in the window the drop occurs and cuts both the read and quality strings there for the 3'-end cut. However, if the length of the remaining sequence is less than the minimum length threshold, then the read is discarded entirely. 5'-end trimming can be disabled.

Sickle also has an option to discard reads with any Ns in them.

Sickle supports three types of quality values: Illumina, Solexa, and Sanger. Note that the Solexa quality setting is an approximation (the actual conversion is a non-linear transformation). The end approximation is close. Illumina quality refers to qualities encoded with the CASAVA pipeline between versions 1.3 and 1.7. Illumina quality using CASAVA >= 1.8 is Sanger encoded. Sickle will get the quality type from the datatype of the file.

Note that Sickle will remove the 2nd fastq record header (on the "+" line) and replace it with simply a "+". This is the default format for CASAVA >= 1.8.

Sickle also supports gzipped file inputs.

Sickle is a sliding window trimmer and tries to keep the longest high quality 5' sequence reads.

windows of N bases moving from 5' to 3' end are tested for average quality. In the first window that fails to meet $>Q$, bases are trimmed starting with the first base with quality $< Q$

FastQ Manipulation

Galaxy / UCLA

Analyze DataWorkflowShared Data ▾Help ▾User ▾

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
- Sickle Windowed Adaptive Trimming of FastQ data
- Scythe Trimming adapters/contaminants using a Naive Bayesian classifier

AB-SOLID DATA

- Quality format converter (ASCII-Numeric)
- Compute quality statistics
- Draw quality score boxplot
- Draw nucleotides distribution chart
- FASTQ to FASTA converter

Scythe (version 1.0.0)

FastQ Reads:

2: FASTQ file ▾

Note: Scythe will infer the quality type of the file from its datatype. I.e., if the datatype is fastqsanger, then the quality type is sanger. The default is fastqsanger.

Adapter/Contaminant file (in fasta format):

▾

Add a tag to the header indicating that Scythe cut a sequence?:

☐

Also output another file with details about adapter/contaminant matches?:

☐

Prior:

0.3

The prior contamination rate

Smallest length adapter/contaminant to consider:

5

Filter sequences less than this length (after trimming):

35

Execute

Scythe uses a Naive Bayesian approach to classify contaminant substrings in sequence reads. It considers quality information, which can make it robust in picking out 3'-end adapters, which often include poor quality bases.

Most next generation sequencing reads have deteriorating quality towards the 3'-end. It's common for a quality-based trimmer to be employed before mapping, assemblies, and analysis to remove these poor quality bases. However, quality-based trimming could remove bases that are helpful in identifying (and removing) 3'-end adapter contaminants. Thus, it is recommended you run Scythe before quality-based trimming, as part of a read quality control pipeline.

The Bayesian approach Scythe uses compares two likelihood models: the probability of seeing the matches in a sequence given contamination, and not given contamination. Given that the read is contaminated, the probability of seeing a certain number of matches and mismatches is a function of the quality of the sequence. Given the read is not contaminated (and is thus assumed to be random sequence), the probability of seeing a certain number of matches and mismatches is chance. The posterior is calculated across both these likelihood models, and the class (contaminated or not contaminated) with the maximum posterior probability is the class selected.

Scythe is an adapter trimmer for Illumina reads that employs a Bayesian model to classify contaminant substrings in reads

FastQ Manipulation

The screenshot shows the Galaxy / UCLA web interface. The top navigation bar includes 'Galaxy / UCLA', 'Analyze Data', 'Workflow', 'Shared Data', 'Help', and 'User'. On the left, a 'Tools' sidebar lists various bioinformatics tools, with 'FASTQ Trimmer' highlighted. The main panel displays the 'FASTQ Trimmer (version 1.0.0)' tool. The tool's configuration section includes a 'FASTQ File' dropdown set to '2: FASTQ file', a 'Define Base Offsets as:' dropdown set to 'Absolute Values', and input fields for 'Offset from 5' end' and 'Offset from 3' end', both set to '0'. There is a checkbox for 'Keep reads with zero length' which is currently unchecked. A blue 'Execute' button is located below the configuration fields. Below the button, a text block explains the tool's function: 'This tool allows you to trim the ends of reads. You can specify either absolute or percent-based offsets. Offsets are calculated, starting at 0, from the respective end to be trimmed. When using the percent-based method, offsets are rounded to the nearest integer. For example, if you have a read of length 36: @Some FASTQ Sanger Read CAATATGTNCTCACTGATAAGTGGATATNAGCNCCA + =@. @;B-?8>CBA@>7@7BBCA4-48%<;?<B@'. And you set absolute offsets of 2 and 9: @Some FASTQ Sanger Read ATATGTNCTCACTGATAAGTGGATA + @. @;B-?8>CBA@>7@7BBCA4-4. Or you set percent offsets of 6% and 20% (corresponds to absolute offsets of 2,7 for a read length of 36):'.

Tools

- 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
- 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
- Sickle Windowed Adaptive Trimming of FastQ data
- Scythe Trimming adapters/contaminants using a Naive Bayesian classifier
- AB-SOLID DATA
- Quality format converter (ASCII-

FASTQ Trimmer (version 1.0.0)

FASTQ File:
2: FASTQ file

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:
0
Values start at 0, increasing from the left

Offset from 3' end:
0
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 calculated, starting at 0, from the respective end to be trimmed. When using the percent-based method, offsets are rounded to the nearest integer.

For example, if you have a read of length 36:

```
@Some FASTQ Sanger Read
CAATATGTNCTCACTGATAAGTGGATATNAGCNCCA
+
=@@. @;B-?8>CBA@>7@7BBCA4-48%<;?<B@
```

And you set absolute offsets of 2 and 9:

```
@Some FASTQ Sanger Read
ATATGTNCTCACTGATAAGTGGATA
+
@. @;B-?8>CBA@>7@7BBCA4-4
```

Or you set percent offsets of 6% and 20% (corresponds to absolute offsets of 2,7 for a read length of 36):

Run FASTQ trimmer with 15 as offset from 5' end and 30 as offset from 3' end, then run FastQC with trimmed reads

Mapping Reads to a Genome

Map with BWA for Illumina (version 1.2.3)

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

Use a built-in index

Select a reference genome:

araTha2

Is this library mate-paired?:

Single-end

FASTQ file:

4: FASTQ Trimmer on data 2

FASTQ with either Sanger-scaled quality values (fastqsanger) or Illumina-scaled quality values (fastqillumina)

BWA settings to use:

Commonly Used

For most mapping needs use Commonly Used settings. If you want full control use Full Parameter List

Suppress the header in the output SAM file:

☐

BWA produces SAM with several lines of header information

Execute

What it does

BWA is a fast light-weighted tool that aligns relatively short sequences (queries) to a sequence database (large), such as the human reference genome. It is developed by Heng Li at the Sanger Institute. Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-60.

Know what you are doing

⚠ There is no such thing (yet) as an automated gearshift in short read mapping. It is all like stick-shift driving in San Francisco. In other words = running this tool with default parameters will probably not give you meaningful results. A way to deal with this is to **understand** the parameters by carefully reading the [documentation](#) and experimenting. Fortunately, Galaxy makes experimenting easy.

Input formats

BWA accepts files in either Sanger FASTQ format (galaxy type *fastqsanger*) or Illumina FASTQ format (galaxy type *fastqillumina*). Use the FASTQ Groomer to prepare your files.

A Note on Built-in Reference Genomes

The default variant for all genomes is "Full", defined as all primary chromosomes (or scaffolds/contigs) including mitochondrial plus associated unmapped, plasmid, and other segments. When only one version of a genome is available in this tool, it represents the default "Full" variant. Some genomes will have more than one variant available. The "Canonical Male" or sometimes simply "Canonical" variant contains the primary chromosomes for a genome. For example a human "Canonical" variant contains chr1-chr22, chrX, chrY, and chrM. The "Canonical Female" variant contains the primary chromosomes excluding chrY.

BWA performs gapped alignments and can be used to detect indels and SNPs. BWA is generally used for DNA projects.

Outputs

The output is in SAM format, and has the following columns:

Column	Description
1 QNAME	Query (pair) NAME
2 FLAG	bitwise FLAG
3 RNAME	Reference sequence NAME
4 POS	1-based leftmost POSITION/coordinate of clipped sequence
5 MAPQ	MAPPING Quality (Phred-scaled)
6 CIGAR	extended CIGAR string
7 MRNM	Mate Reference sequence NaMe ('-' if same as RNAME)
8 MPOS	1-based Mate POSITION
9 ISIZE	Inferred insert SIZE
10 SEQ	query SEQUENCE on the same strand as the reference
11 QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12 OPT	variable OPTional fields in the format TAG:VTYPE:VALU

The flags are as follows:

Flag	Description
0x0001	the read is paired in sequencing
0x0002	the read is mapped in a proper pair
0x0004	the query sequence itself is unmapped
0x0008	the mate is unmapped
0x0010	strand of the query (1 for reverse)
0x0020	strand of the mate
0x0040	the read is the first read in a pair
0x0080	the read is the second read in a pair
0x0100	the alignment is not primary

It looks like this (scroll sideways to see the entire example):

QNAME	FLAG	RNAME	POS	MAPQ	CIGAR	MRNM	MPOS	ISIZE	SEQ	QUAL	OPT	
HWI-EAS91_1_30788AAXX:1:1:1761:343	4	*	0	0	*	*	0	0	AAAAAAN			
HWI-EAS91_1_30788AAXX:1:1:1578:331	4	*	0	0	*	*	0	0	GTATAGAN			

BWA settings

All of the options have a default value. You can change any of them. All of the options in BWA have been implemented here.

BWA parameter list

This is an exhaustive list of BWA options:

For **aln**:

- n NUM Maximum edit distance if the value is INT, or the fraction of missing alignments given 2% uniform base error rate if FLOAT. In the latter case, the maximum edit distance is automatically chosen for different read lengths. [0.04]
- o INT Maximum number of gap opens [1]
- e INT Maximum number of gap extensions, -1 for k-difference mode (disallowing long gaps) [-1]
- d INT Disallow a long deletion within INT bp towards the 3'-end [16]

RNA-Seq Aligners

- ✓ Bowtie
 - It doesn't perform gapped alignments. It runs faster and requires smaller memory footprint.
 - ✓ Bowtie2
 - It is fast and can perform local and gapped alignment. It performs better for reads longer than 50bp.
- Bowtie and bowtie2 use indexed reference genome
- ✓ Tophat
 - Most popular splice junction mapper for RNA-Seq reads. It first uses bowtie to align reads, and then analyzes the mapping reads to identify splice junctions between exons.

Bowtie for RNA-Seq

Map with Bowtie for Illumina (version 1.1.2)

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

Built-ins were indexed using default options

Select a reference genome:

if your genome of interest is not listed – contact Galaxy team

Is this library mate-paired?:

FASTQ file:

Must have ASCII encoded quality scores

Bowtie settings to use:

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

Skip the first n reads (-s):

Only align the first n reads (-u):

-1 for off

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

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

Maximum number of mismatches permitted in the seed (-n):

May be 0, 1, 2, or 3

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

Seed length (-l):

Minimum value is 5

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

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

-1 for default MAQ-like alignment policy

Whether or not to try as hard as possible to find valid alignments when they exist (-y):

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

Seed length (-l):

Minimum value is 5

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

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

-1 for default MAQ-like alignment policy

Whether or not to try as hard as possible to find valid alignments when they exist (-y):

Tryhard mode is much slower than regular mode

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

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

Suppress all alignments for a read if more than n reportable alignments exist (-m):

-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 total mismatches (-b):

Removes all strand bias. Only affects which alignments are reported by Bowtie. Runs slower with best option

Maximum number of backtracks permitted when aligning a read (--maxbts):

Override the offset of the index to n (-o):

-1 for default

Seed for pseudo-random number generator (--seed):

-1 for default

Suppress the header in the output SAM file:

☐

Bowtie produces SAM with several lines of header information by default

Select 'mm10' as reference genome
Select trimmed reads as input for FASTQ file
Change Suppress all alignments for a read to 1 (-m 1)

Sequence Alignment/Map Format (SAM)

- ✓ A generic nucleotide alignment format that describes the alignment of reads to a reference genome in text format.
- ✓ It consists of optional header section and alignment section.

```
Coord      12345678901234  5678901234567890123456789012345
ref        AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGCGCCAT

+r001/1      TTAGATAAAGGATA*CTG
+r002        aaaAGATAA*GGATA
+r003        gcctaAGCTAA
+r004                ATAGCT.....TCAGC
-r003                ttagctTAGGC
-r001/2                CAGCGGCAT
```

The corresponding SAM format is:

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001  99 ref  7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *
r002   0 ref  9 30 3S6M1P1I4M * 0 0 AAAAGATAAAGGATA *
r003   0 ref  9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;
r004   0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```

Alignment Summary

- ✓ Best if more than 80% reads aligned to the reference
- ✓ good library if 60% aligned
- ✓ less than 20%, not complete reference or sample contamination

Picard – SAM/BAM Alignment Summary Metrics

SAM/BAM Alignment Summary Metrics (version 1.56.0)

SAM/BAM dataset to generate statistics for:

11: Map with Bowtie f...apped reads

If empty, upload or import a SAM/BAM dataset.

Title for the output file:

Picard Alignment Summary Metrics

Use this remind you what the job was for.

Select Reference Genome:

Use the assigned data genome/build

Check the assigned reference genome:

Mouse: mm10

Galaxy thinks that the reads in you dataset were aligned against this reference. If this is not appropriate Reference.

Assume the input file is already sorted:



Input file contains Bisulphite sequenced reads:



Adapter sequences:

One per line if multiple

Larger paired end reads and inter-chromosomal pairs considered chimeric:

100000

Execute

```
## net.sf.picard.metrics.StringHeader
```

```
# Started on: Tue Feb 03 17:13:21 PST 2015
```

```
## METRICS CLASS net.sf.picard.analysis.AlignmentSummaryMetrics
```

CATEGORY	UNPAIRED
TOTAL_READS	600000
PF_READS	600000
PCT_PF_READS	1
PF_NOISE_READS	0
PF_READS_ALIGNED	72186
PCT_PF_READS_ALIGNED	0.12031
PF_ALIGNED_BASES	3970230
PF_HQ_ALIGNED_READS	72186
PF_HQ_ALIGNED_BASES	3970230
PF_HQ_ALIGNED_Q20_BASES	3487209
PF_HQ_MEDIAN_MISMATCHES	0
PF_MISMATCH_RATE	0.017576
PF_HQ_ERROR_RATE	0.017576
PF_INDEL_RATE	0
MEAN_READ_LENGTH	55
READS_ALIGNED_IN_PAIRS	0
PCT_READS_ALIGNED_IN_PAIRS	0
BAD_CYCLES	0
STRAND_BALANCE	0.500776
PCT_CHIMERAS	0
PCT_ADAPTER	0.002177
SAMPLE	
LIBRARY	
READ_GROUP	

Picard Tool Run Log

```
INFO:root:## executing samtools view -h -b -S -o /u/home/galaxy/galaxy/galaxy-dist/database/job_working_directory/061/61959/dataset_112721
```

```
INFO:root:## executing samtools sort /u/home/galaxy/galaxy/galaxy-dist/database/job_working_directory/061/61959/dataset_112721_files/tmpNDc
```

```
INFO:root:## executing java -Xmx4g -Djava.io.tmpdir="/u/home/galaxy/galaxy/galaxy-dist/database/tmp" -jar /u/home/galaxy/galaxy/galaxy-di
```

Uncheck “assume the input file is already sorted”

Extract Workflow

Workflow name

Workflow constructed from history 'qc'

Create Workflow

Check all

Uncheck all

Tool	History items created
Upload File <i>This tool cannot be used in workflows</i>	1: s_1_1_600000_qseq.txt ☒ Treat as input dataset
qseq_to_fastq ☒ Include "qseq_to_fastq" in workflow	2: FASTQ file
Fastqc: Fastqc QC ☒ Include "Fastqc: Fastqc QC" in workflow	3: FastQC_data 2.html
Compute quality statistics ☒ Include "Compute quality statistics" in workflow	5: Compute quality statistics on data 2
Draw quality score boxplot ☒ Include "Draw quality score boxplot" in workflow	7: Draw quality score boxplot on data 5
Draw nucleotides distribution chart	8: Draw nucleotides distribution chart on data

History

qc

HISTORY LISTS

Saved Histories

Histories Shared with Me

CURRENT HISTORY

Create New

Clone

Copy Datasets

Share or Publish

Extract Workflow

Dataset Security

Show Deleted Datasets

Show Hidden Datasets

Purge Deleted Datasets

Show Structure

Export to File

Delete

Delete Permanently

OTHER ACTIONS

Import from File

14: Picard Summary

11: Map v Illumina c

10: FastQ

9: FASTQ data 2

8: Draw n distributi

7: Draw q boxplot o

5: Compu statistics

3: FastQC

2: FASTQ file

142.1 Mb

format: fastqsanger, database: mm9

Workflow Management

Galaxy / UCLA Analyze Data Workflow

Your workflows

Name
qc-bowtie-pic
ceas
peak calling

Workflows by others

Name	Owner
demultiplex	mygalaxyworkshop_1@galaxy.ucla.edu

Other options

Configure your workflow menu

- Edit
- Run
- Share or Publish
- Download or Export
- Clone
- Rename
- View
- Delete

Published workflow/history listed as shared data

A new workflow can be created from scratch or import from a published workflow

Workflow Canvas | peak calling

Options: Save, Run, Edit Attributes, Auto Re-layout, Close

Details: Map with Bowtie for

select a reference from your history or use a built-in index?

Use a built-in index

Select a reference genome: ▼ araTha2

Is this library mate-paired?: Single-end

FASTQ file

Data input 'sinput1' (fastqsanger or fastqillumina or fastqsolexa)

Bowtie settings to use: Full parameter list

Skip the first n reads (-s): ▼ 0

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

Input dataset

output

MACS14

ChIP-Seq Tag File

ChIP-Seq Control File

output_bed_file (bed)

output_xls_to_interval_peaks_file (interval)

output_xls_to_interval_negative_peaks_file (interval)

output_treatment_wig_file (wig)

output_control_wig_file (wig)

output_extra_files (html)

Map with Bowtie for Illumina

FASTQ file

output (sam)

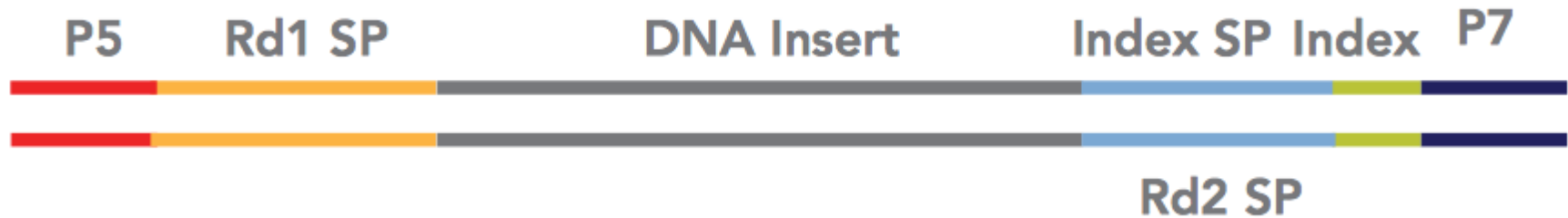
output_suppressed_reads_l (fastq)

output_suppressed_reads_r (fastq)

output_unmapped_reads_l (fastq)

output_unmapped_reads_r (fastq)

Multiplex Sequencing



- ✓ During library preparation, adapters are ligated to the DNA fragments.
 - Rd1 SP and Rd2 SP: primer sites
 - Index SP: primer site for the index read
 - P5 and P7: flow cell attachment sites
- ✓ Index (barcode) allows for sample identification
- ✓ Increase experimental scalability while reduce time and cost
- ✓ Attenuate lane effects

Demultiplexing of FastQ Sequences

- ✓ Barcode splitter

It splits the FastQ data with barcode included in 5' or 3' end of sequence reads.

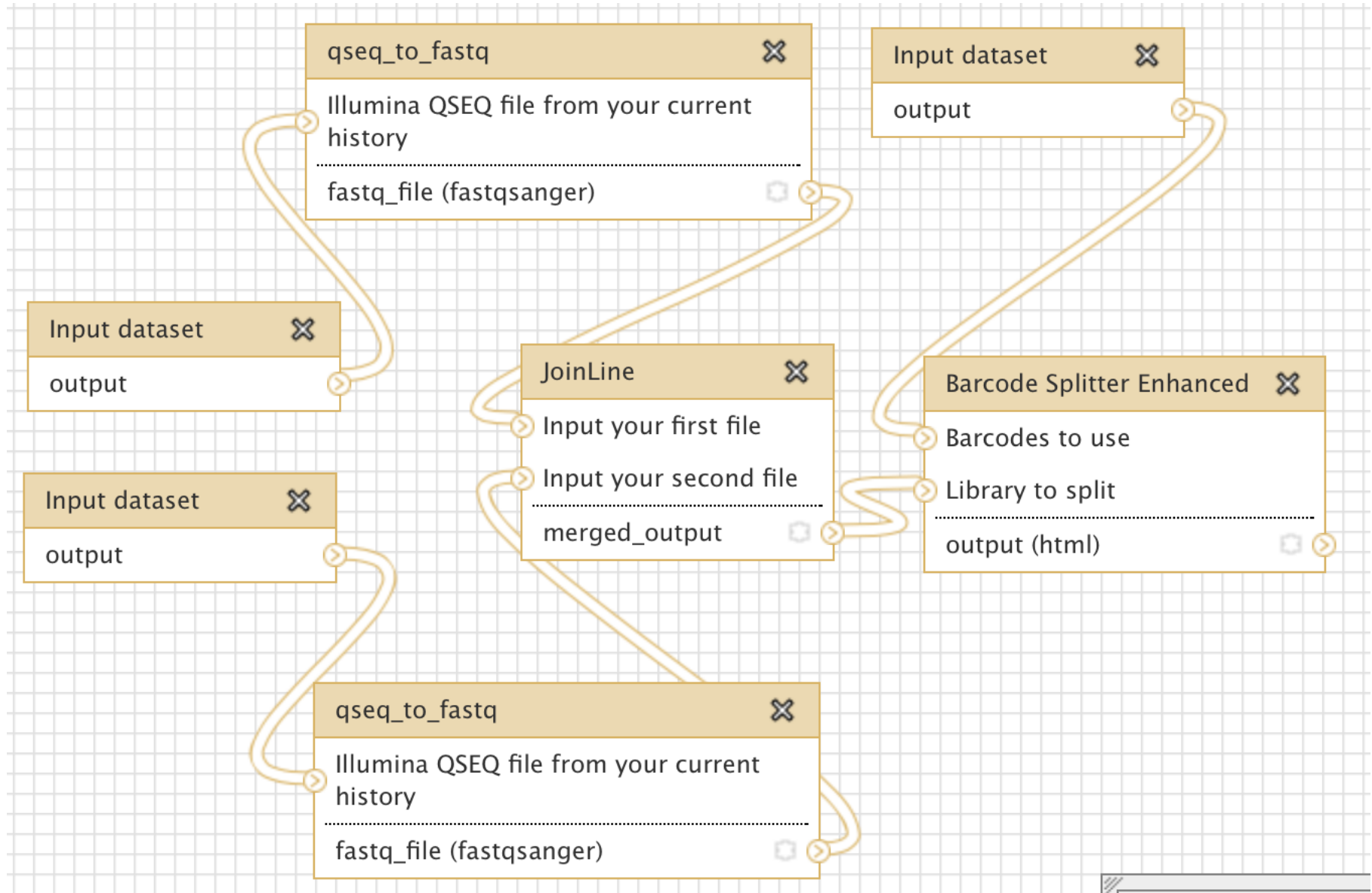
- ✓ Demultiplex workflow

The workflow perform demultiplexing of FastQ sequence data with barcodes and sequences in two separate files.

```
@2:1101:1074.60:113.50:Y
CGATGTA
+2:1101:1074.60:113.50:Y
@@@FDDD
@2:1101:1065.90:113.60:N
CAGATCA
+2:1101:1065.90:113.60:N
CCCFFFF
@2:1101:1067.40:113.90:N
CAGATCA
+2:1101:1067.40:113.90:N
CC@FFFF
```

```
@2:1101:1074.60:113.50:Y
CAGCTCATTGATGCAGTCCAGGCACTTCCCCACATCTCTTCATGTAGGT
+2:1101:1074.60:113.50:Y
;<<DDB:AF?D<FGGBGEGCEE3:CGIII3CC:CC:DFE<D9???9?FDFA
@2:1101:1065.90:113.60:N
CAAAGGGGGGCTTGGTGGGGTGGTTCACGCCTGTAATCCCAGCACTCTGG
+2:1101:1065.90:113.60:N
:>7<?A7A<;::5:(+())&&)0:28(3:78:::33(+++88++8(83(+
@2:1101:1067.40:113.90:N
GATTAGGGTGCTTAGCTGTAACTAAGCGTTTGTGGGTTTAAGTCCCATG
+2:1101:1067.40:113.90:N
+?<D+A:B2<?DDDEBB9?:<CEE4+2+<1):C??@<08BD?D*?*)
```

Demultiplexing workflow



Demultiplexing of FastQ Sequences

- ✓ Upload s_2_2_1101_cut_qseq.txt, s_2_1_1101_cut_qseq.txt, barcode.txt to galaxy
- ✓ Convert qseq files to fastq files
- ✓ Run JoinLine program
- ✓ Run barcode splitter enhanced program
- ✓ Rename dataset to match sample name
- ✓ Run QC workflow for the splitted sample sequence datasets as needed.