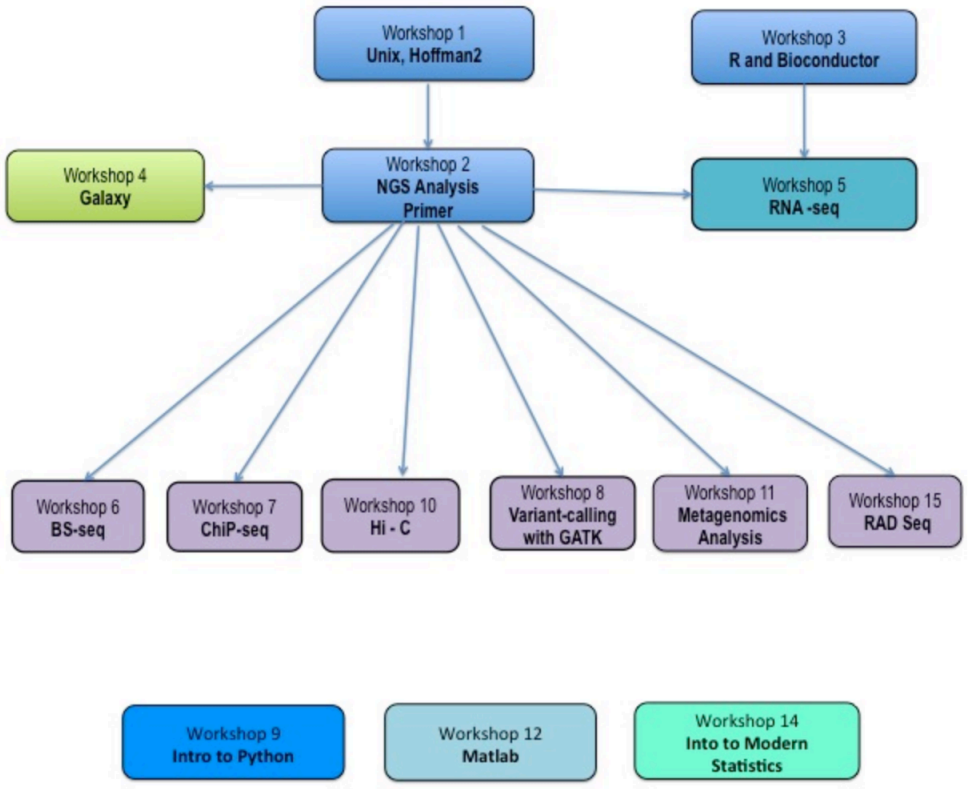
Galaxy Platform For NGS Data Analyses

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

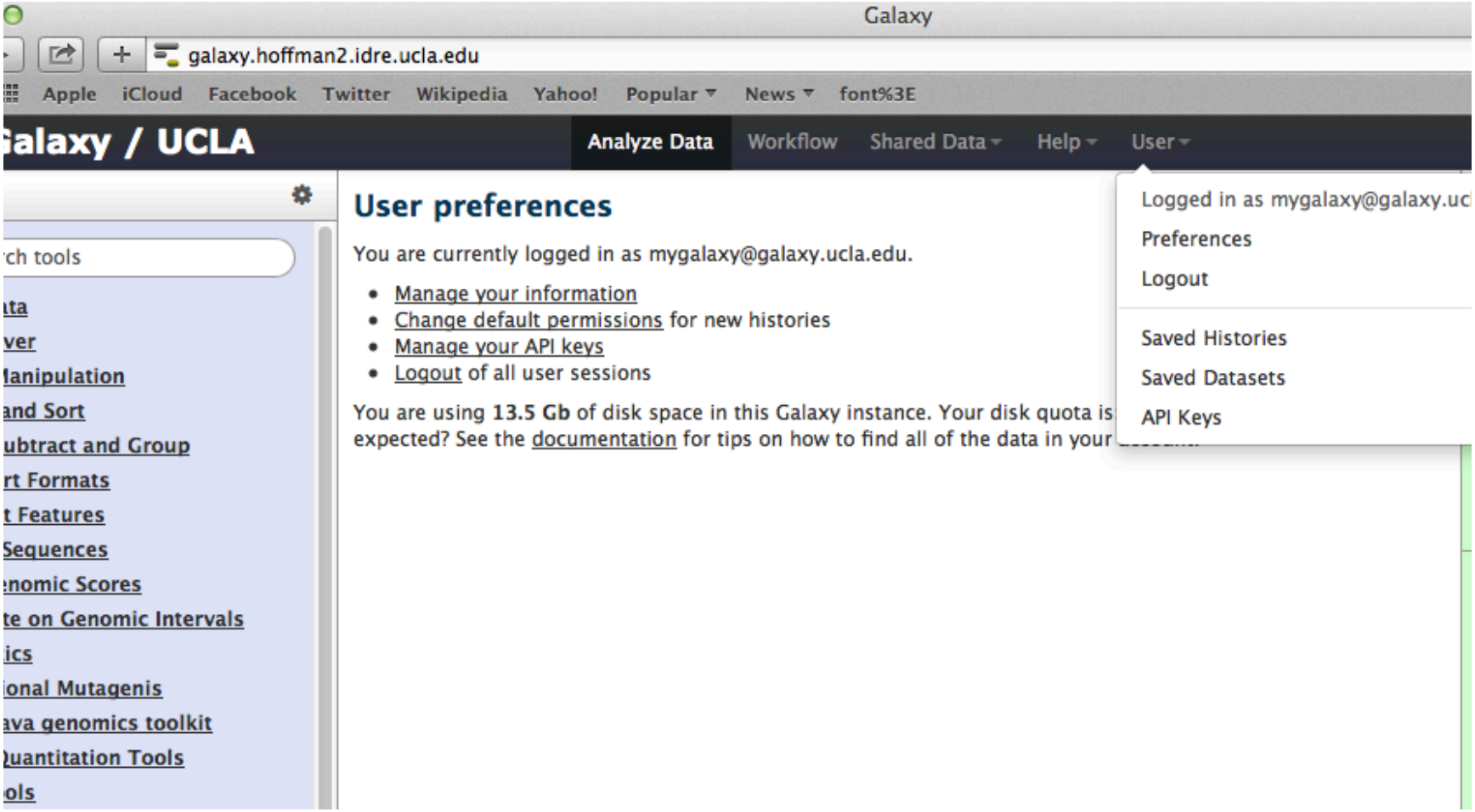
- ✓ galaxy login account:

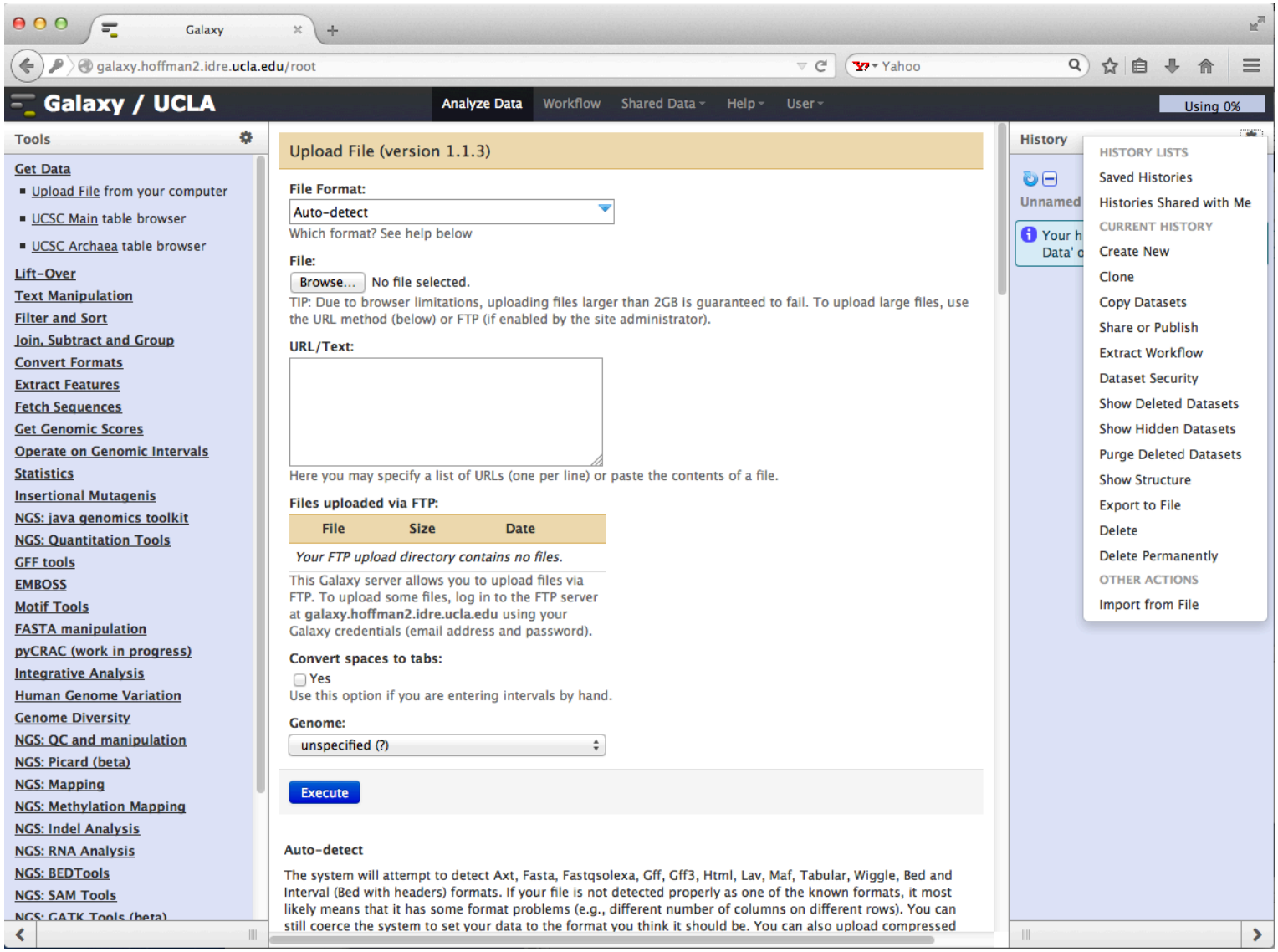
  login: your email associated with ucla

- ✓ Disk quota:

  250Gb/user

Galaxy Account Management

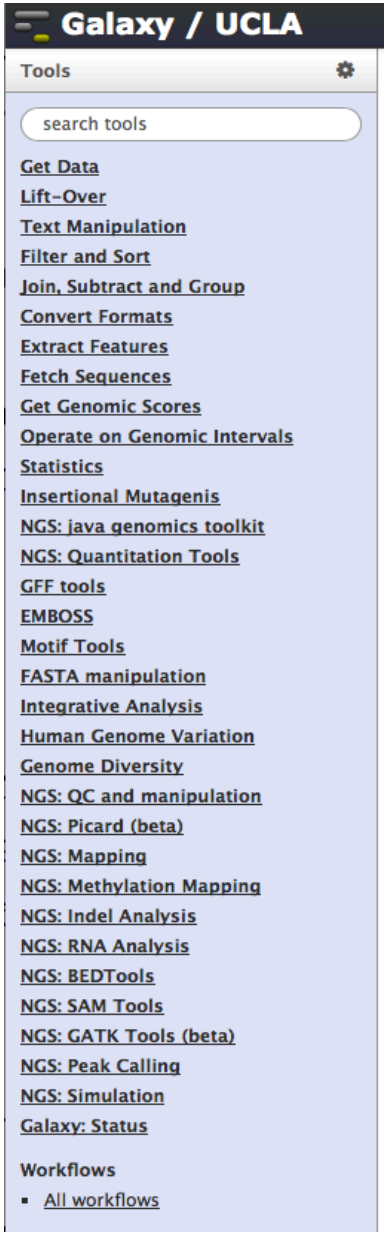


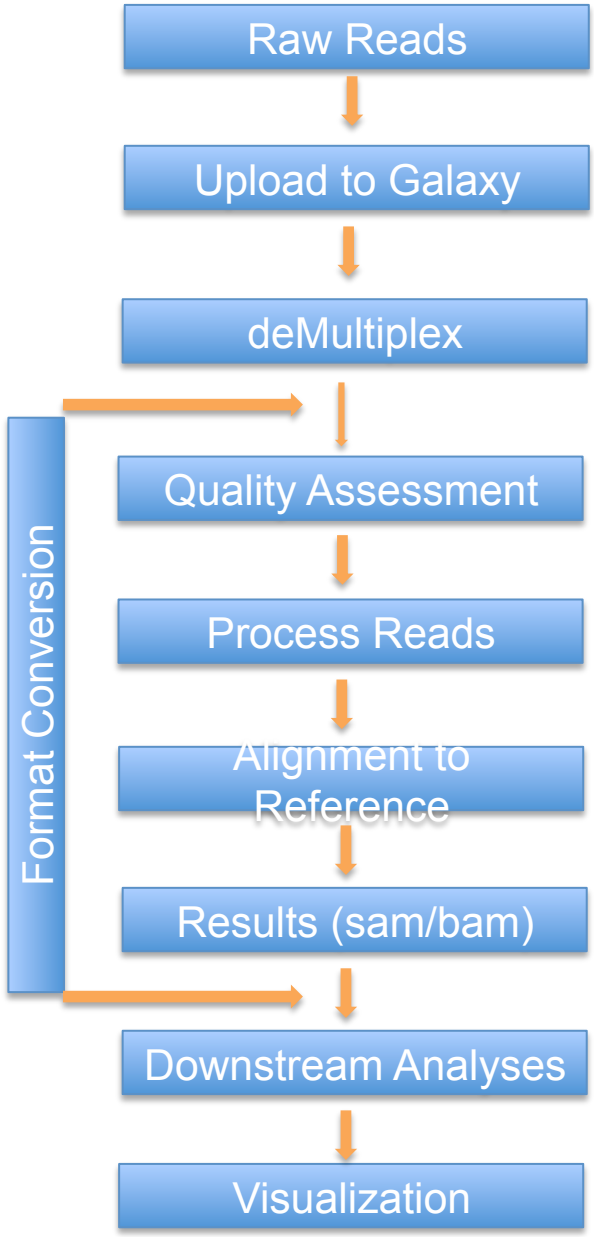


Installed tools

Launch analysis and view result

History of execution and results





*_qseq.txt, *.fastq

File transfer protocol (ftp)

Barcode splitter, deMultiplex workflow

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

Trim sequences, sickle, scythe

bwa, bowtie, bowtie2, tophat

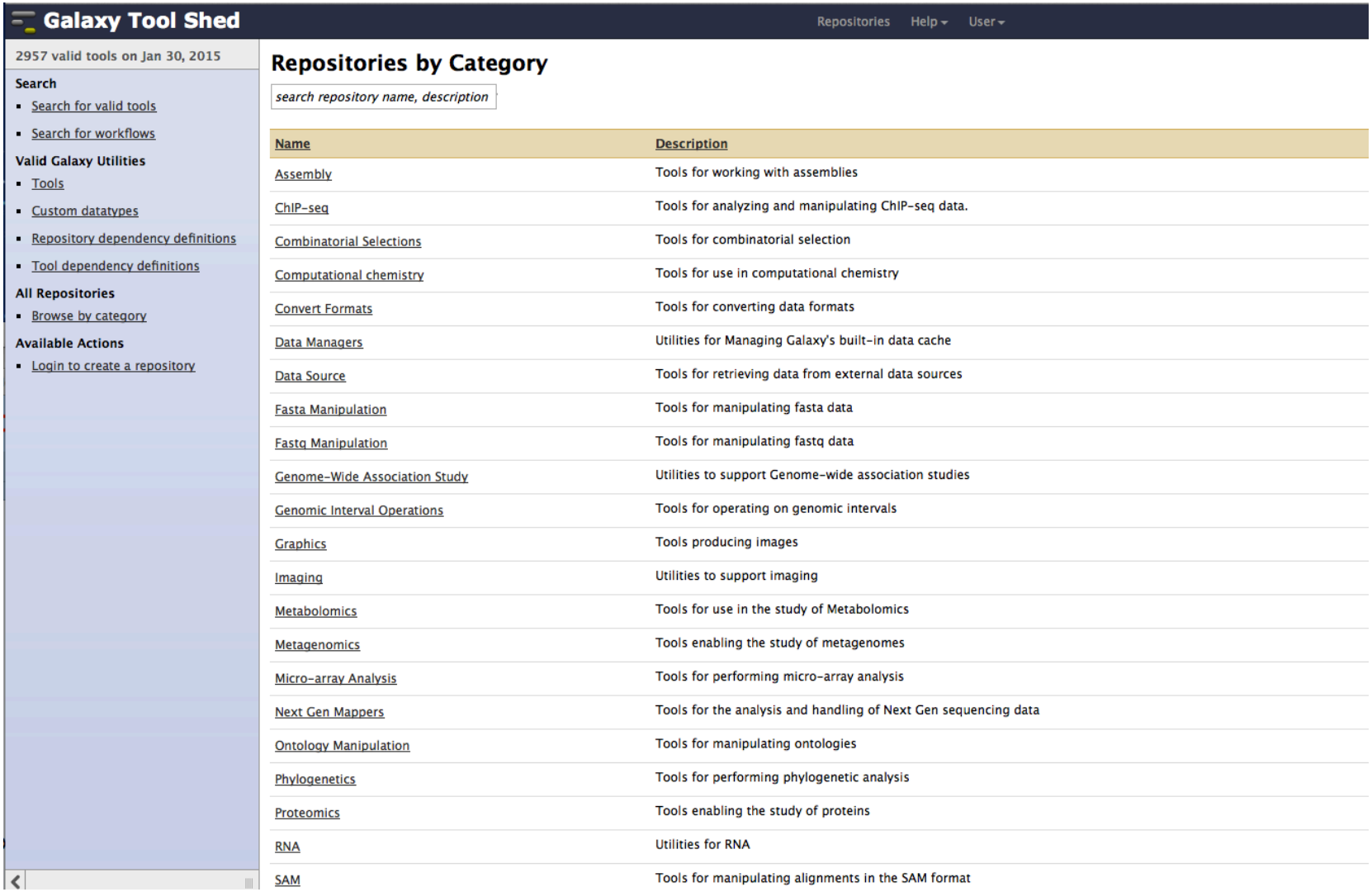
Text manipulation toolkit, BEDTools, SAM Tools, java genomics toolkit, picard toolkit

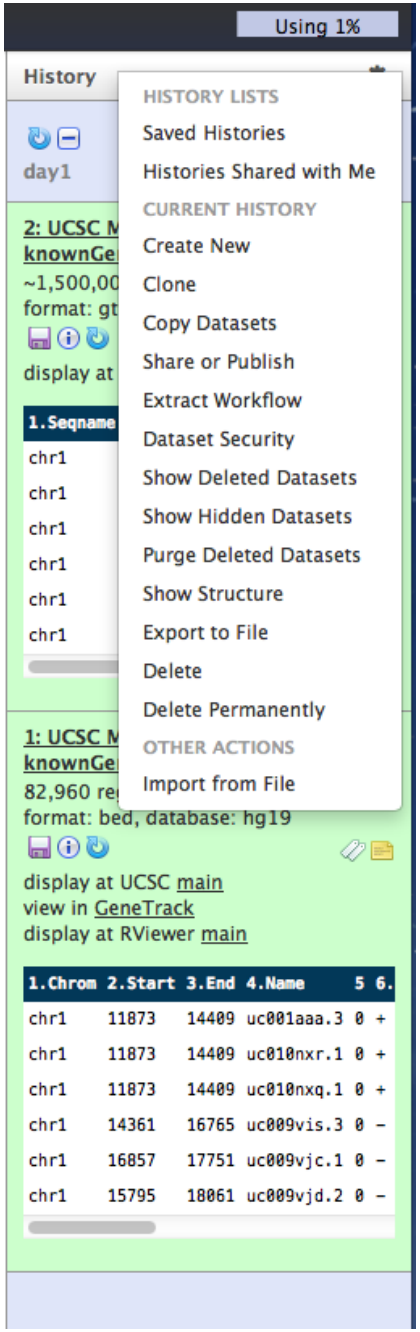
BS-Seeker2, cufflinks, cuffdiff, macs, macs2, GATK, CEAS

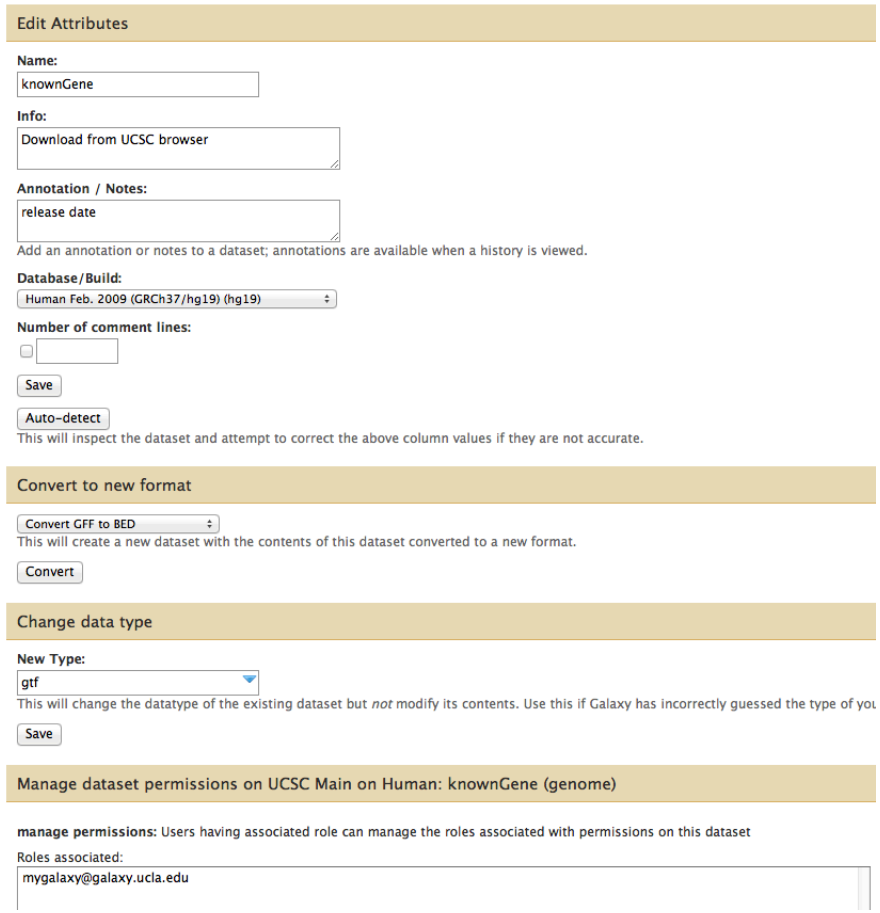
Genome browser, IGV

Repositories of Galaxy Tools

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

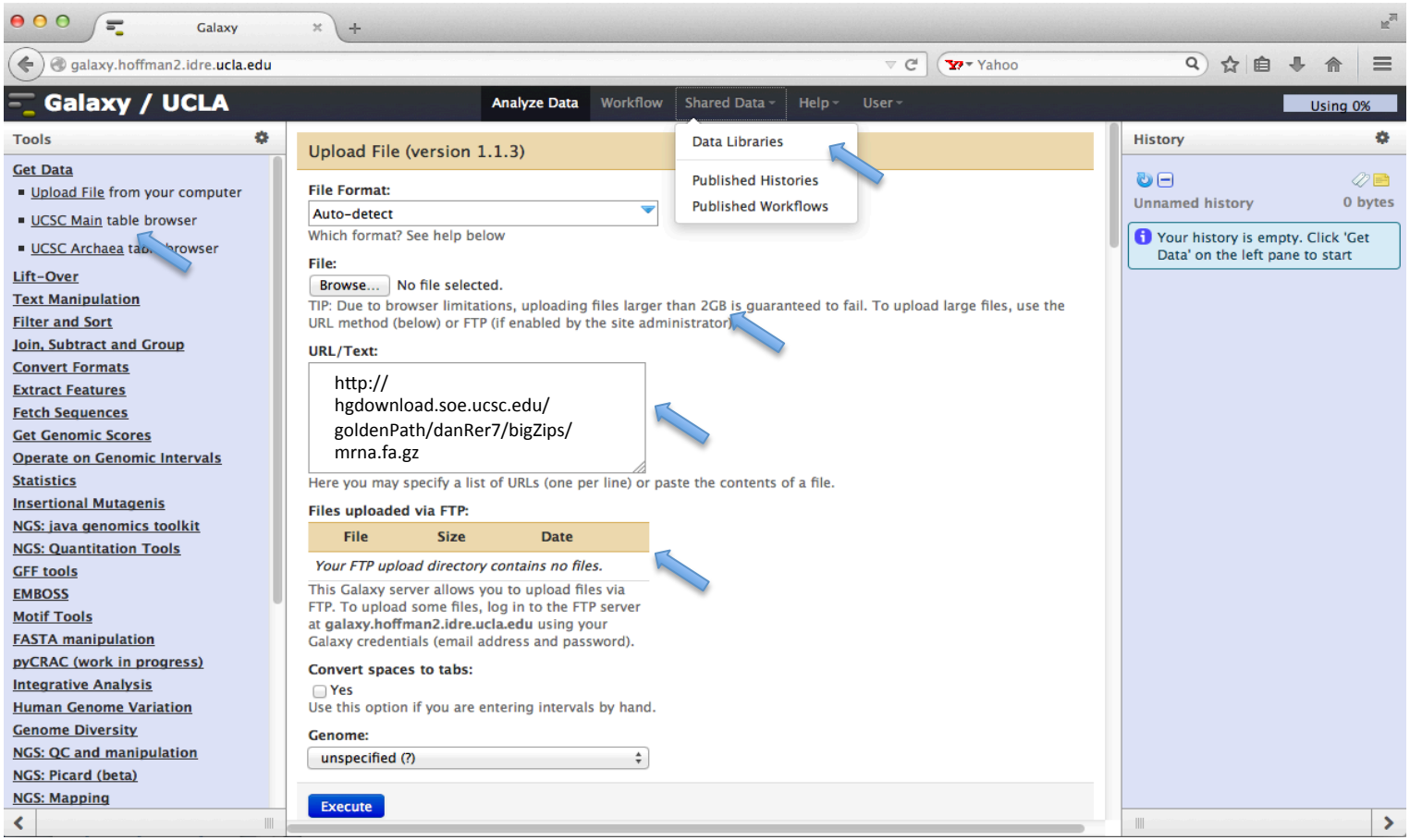






- ✓ History panel contains all datasets that are uploaded and results derived from certain analyses
- ✓ A history can be organized, annotated, and managed as a project
- ✓ History is sharable.
- ✓ Workflow is extracted and built from a history
- ✓ Each dataset under a history can be viewed, examined,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

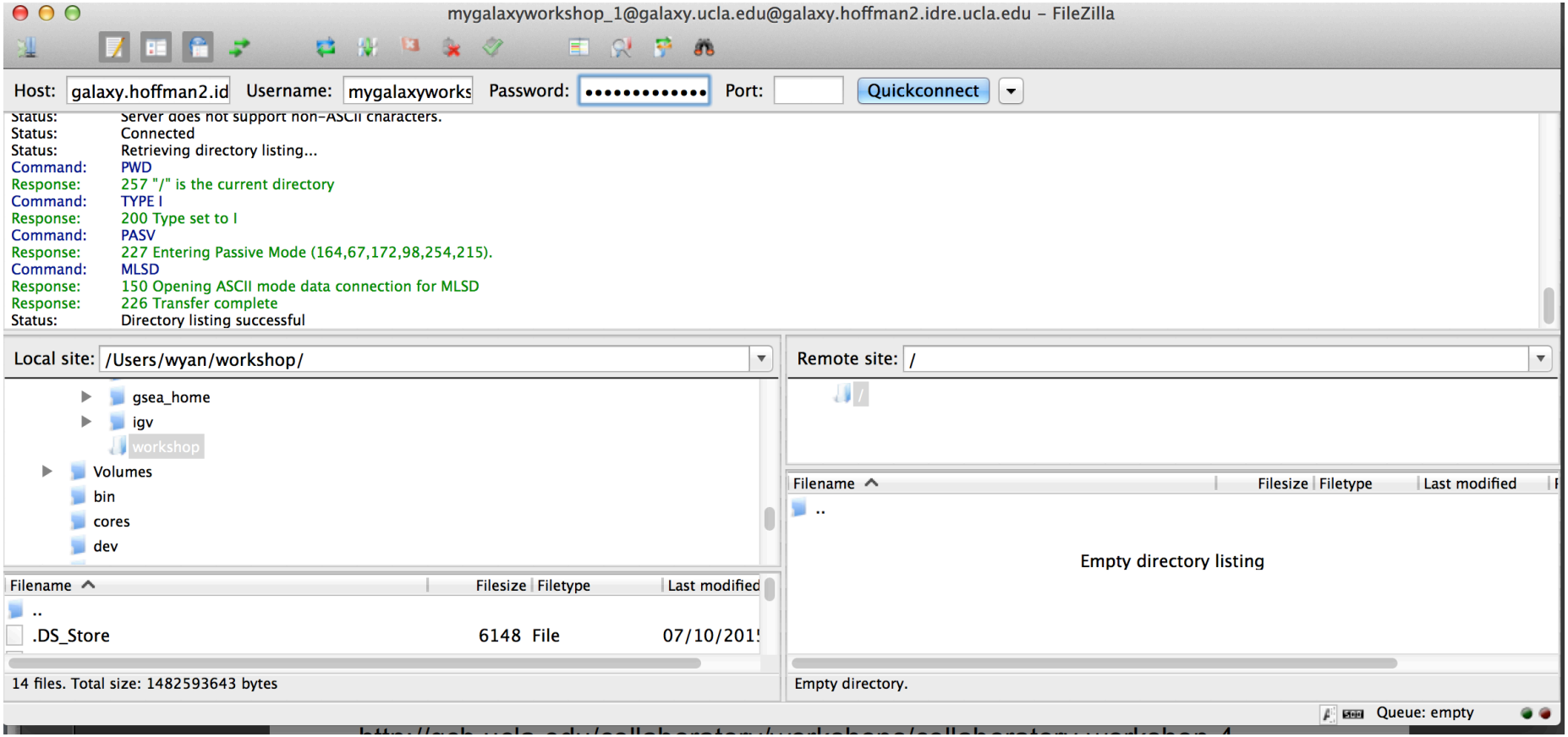
Getting Data to the Galaxy



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

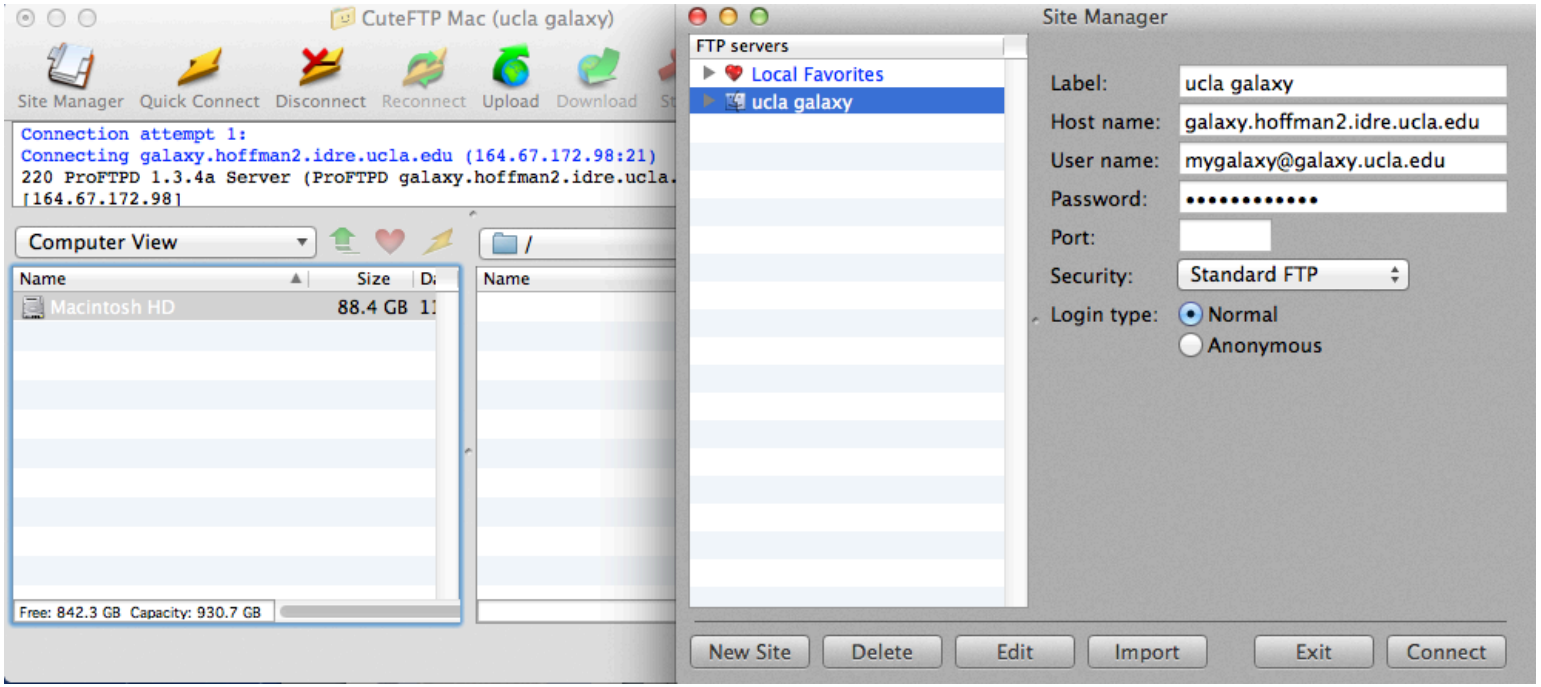
Data libraries: datasets need to be put on the galaxy server before they can be uploaded.

(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/

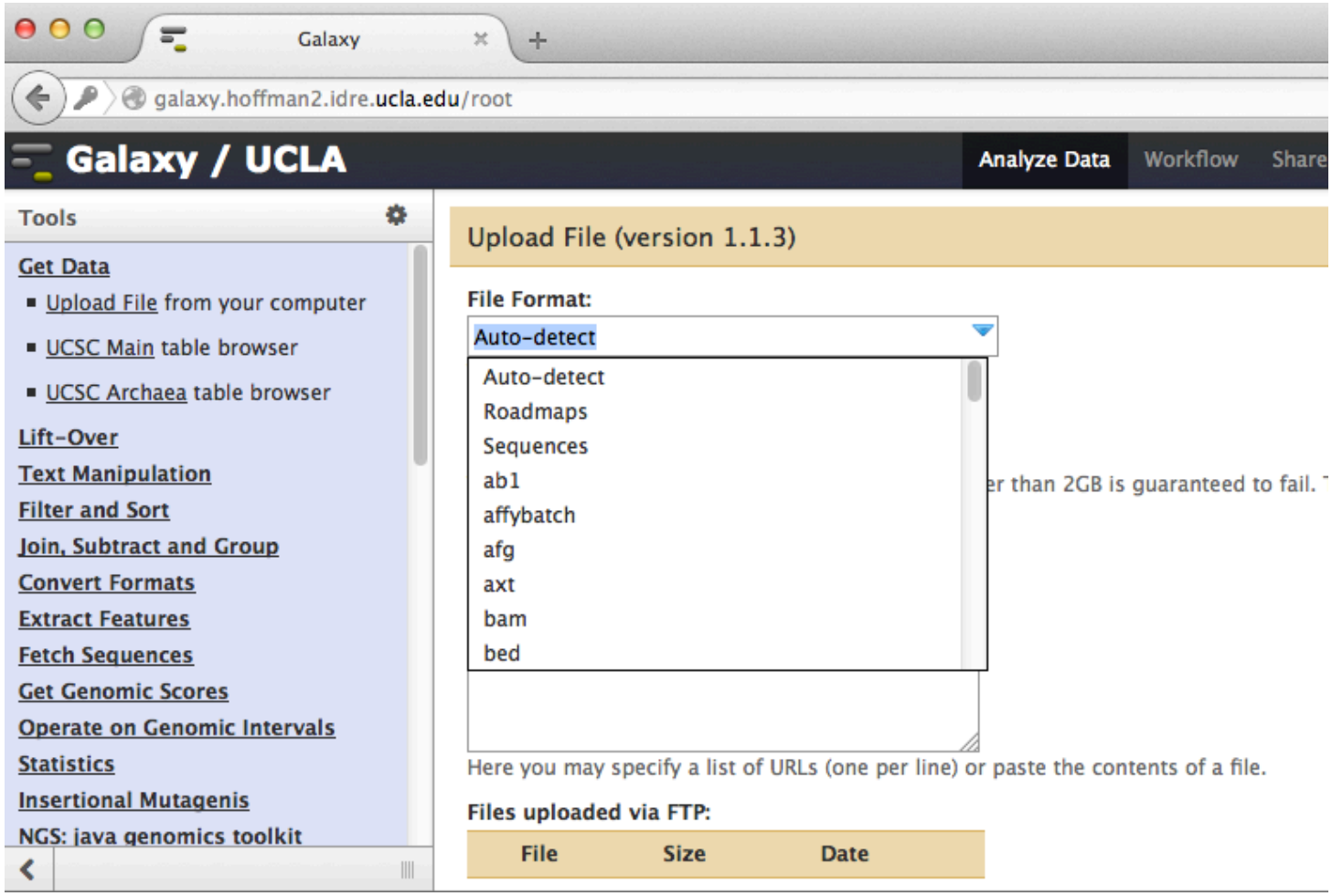
Upload Data to Galaxy

- ✓ Data Transfer from your Hoffman2 Account”

```
[wyan@login3 ~]$ module load gftp
[wyan@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



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

genome.ucsc.edu/FAQ/FAQformat.html

Apple iCloud Facebook Twitter Wikipedia Yahoo! Popular

Frequently Asked Questions: Data File Formats

General formats:

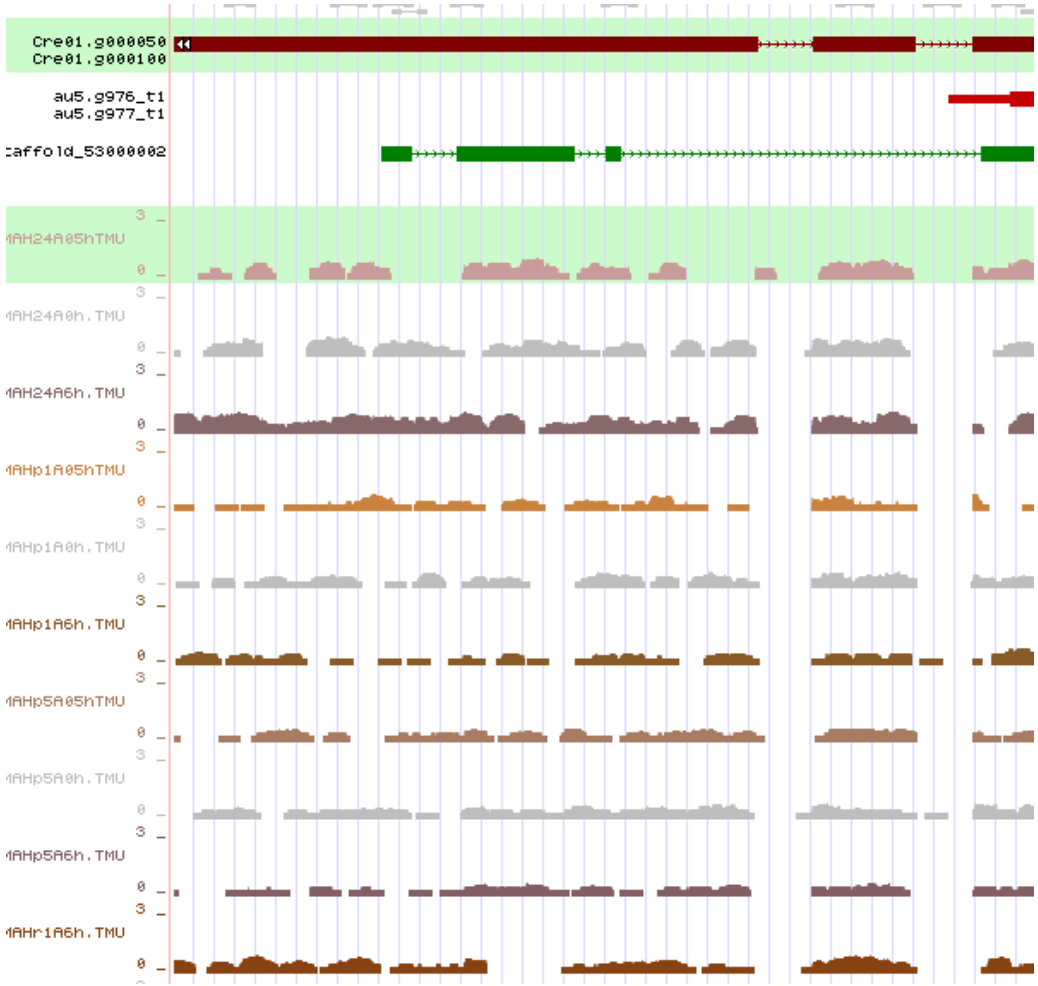
- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- bigBed format
- bigWig format
- Chain format
- GenePred table format
- GFF format
- GTF format
- HAL format
- MAF format
- Microarray format
- Net format
- Personal Genome SNP format
- PSL format
- VCF format
- WIG format

ENCODE-specific formats:

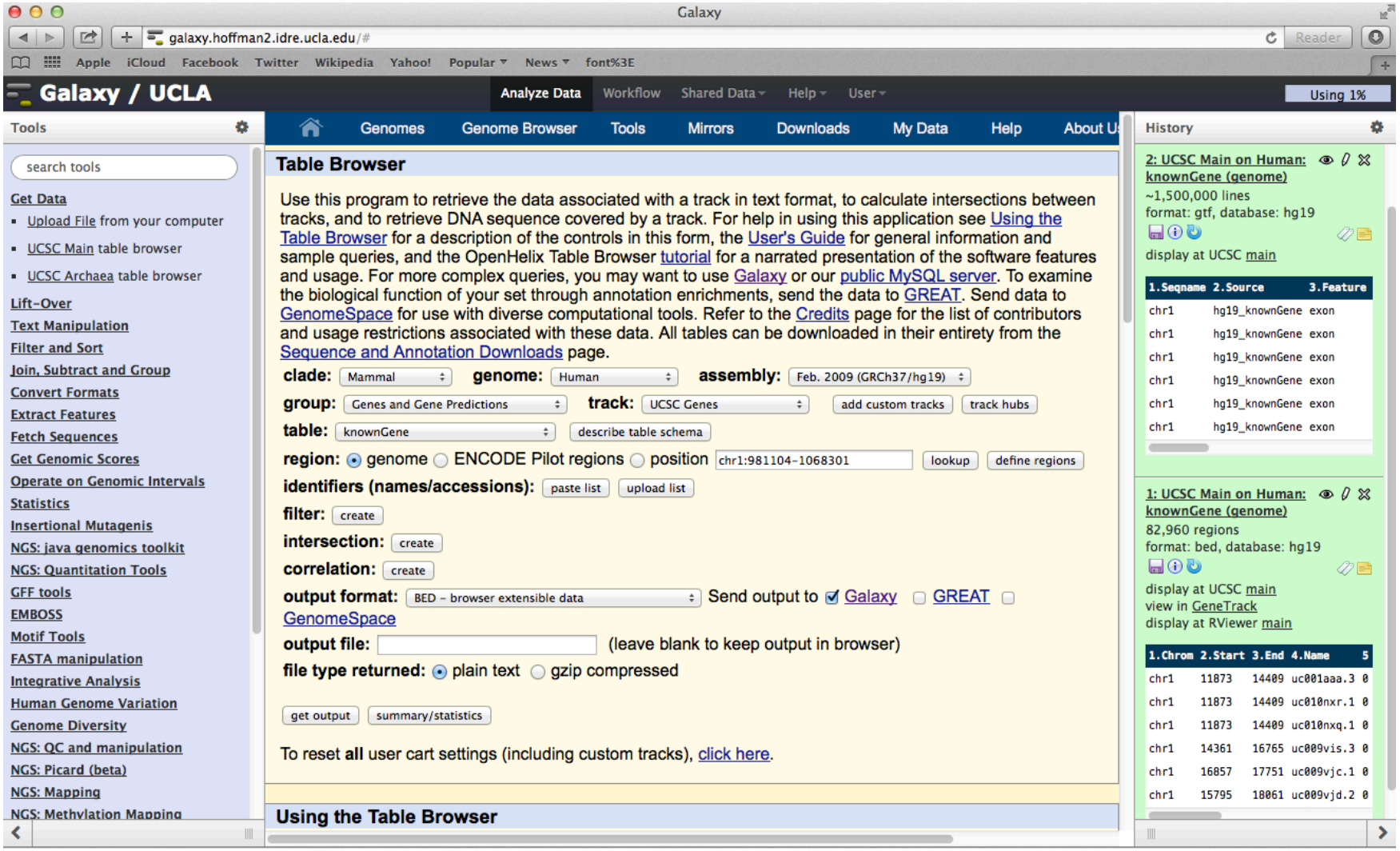
- ENCODE broadPeak format
- ENCODE gappedPeak format
- ENCODE narrowPeak format
- ENCODE pairedTagAlign format
- ENCODE peptideMapping format
- ENCODE RNA elements format
- ENCODE tagAlign format

Download only formats:

- .2bit format
- .fasta format
- .fastQ format
- .nib format



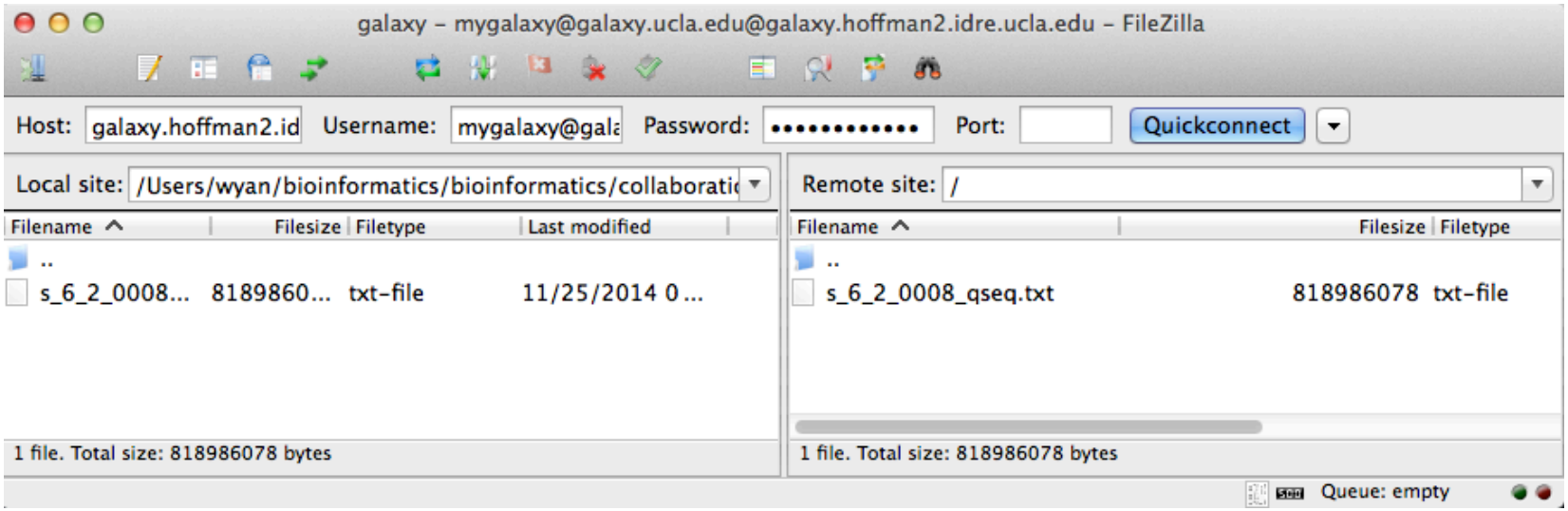
Retrieving Data from UCSC



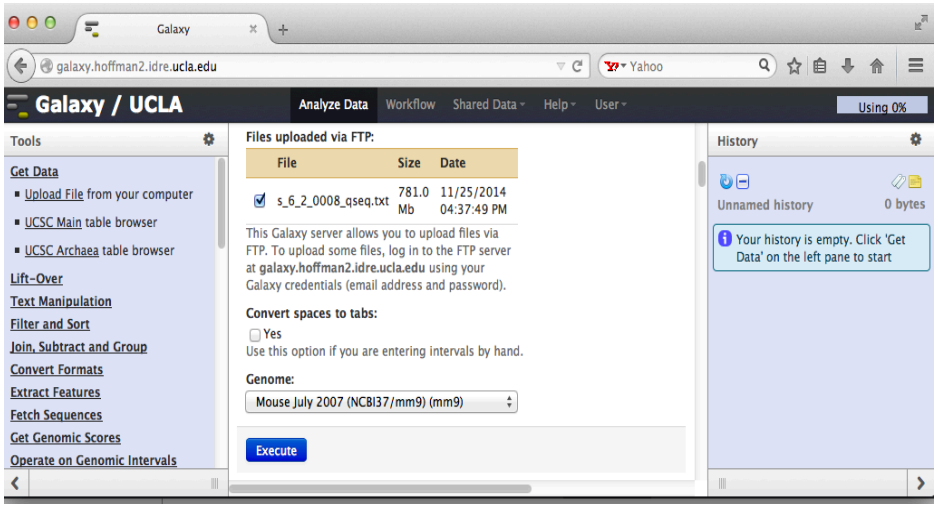
Retrieve knownGene table in two formats from UCSC genome site

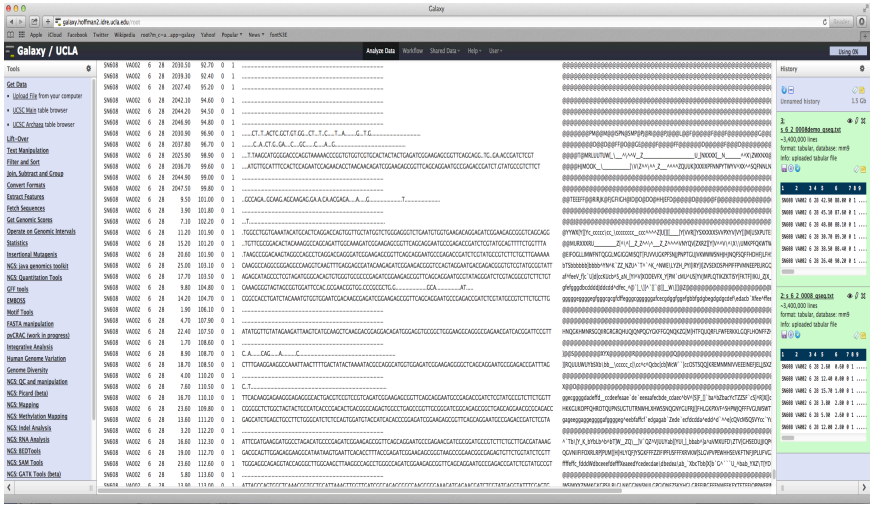
Genomes Pre-installed in Galaxy

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
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

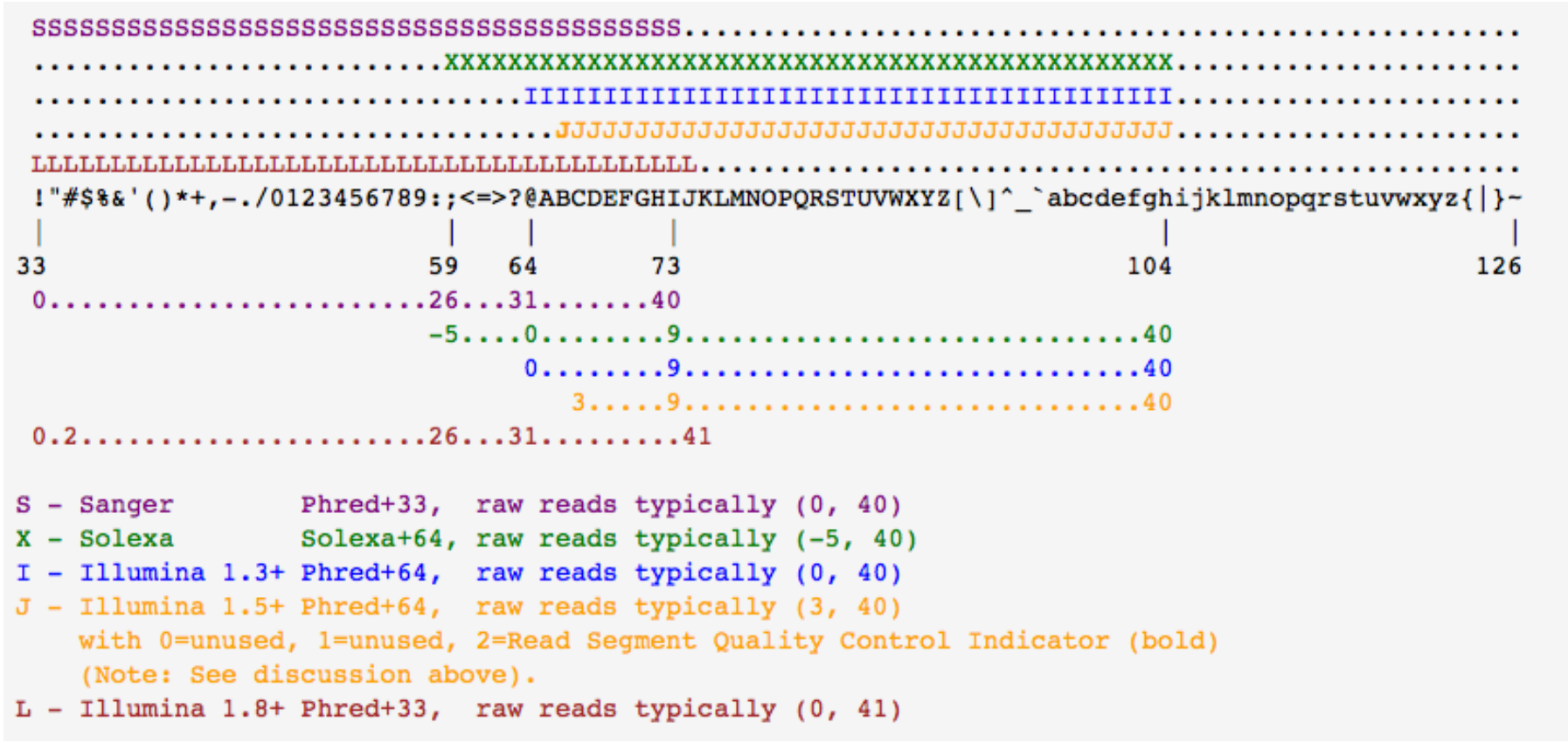
The character '!' represents the lowest quality while '~' is the highest. Here are the quality value characters in left-to-right increasing order of quality (ASCII):

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

FASTQ files from the NCBI/EBI Sequence Read Archive often include a description, e.g.

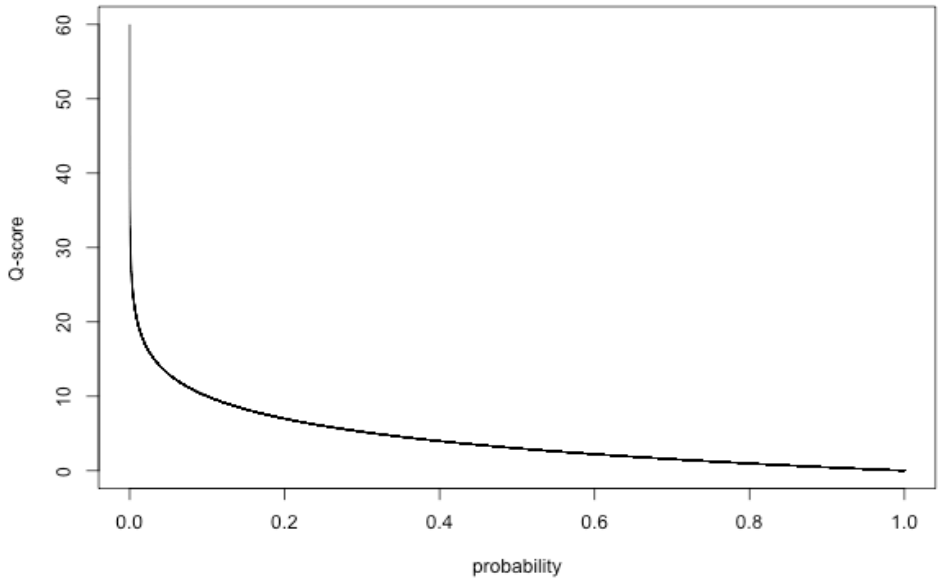
```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

FastQ Quality Scores



Phred Quality Score

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



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.

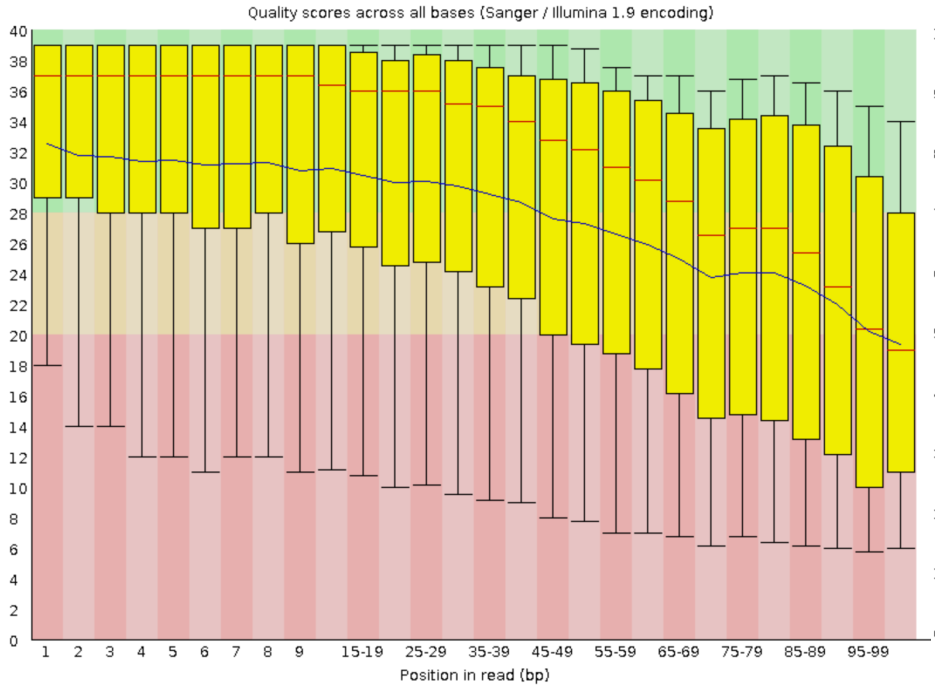
  Trim reads with lower quality score.

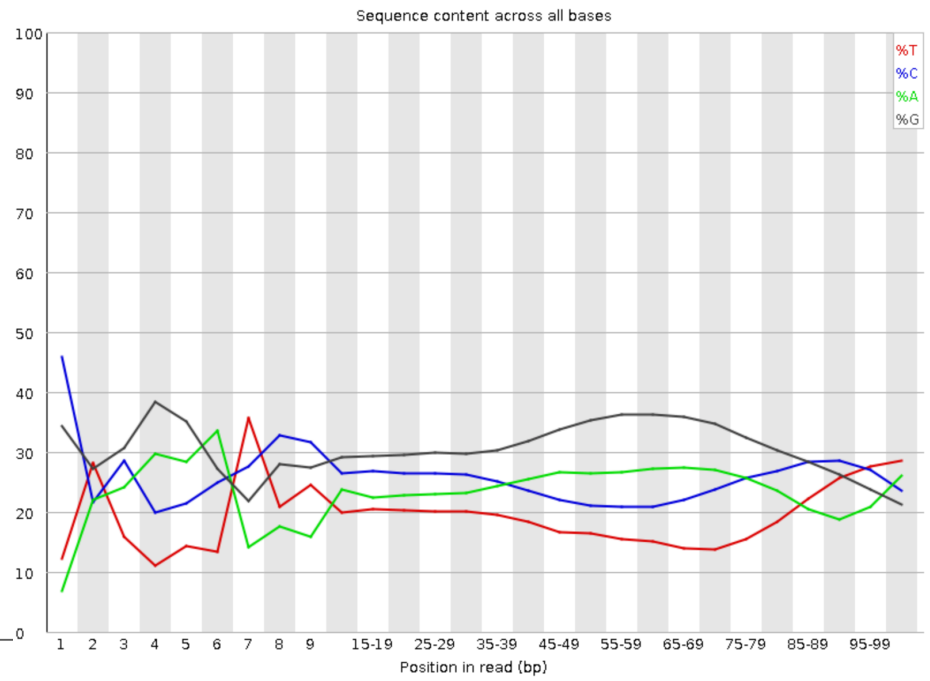
- ✓ Per base sequence and GC content

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

Quality Control of Raw Sequences

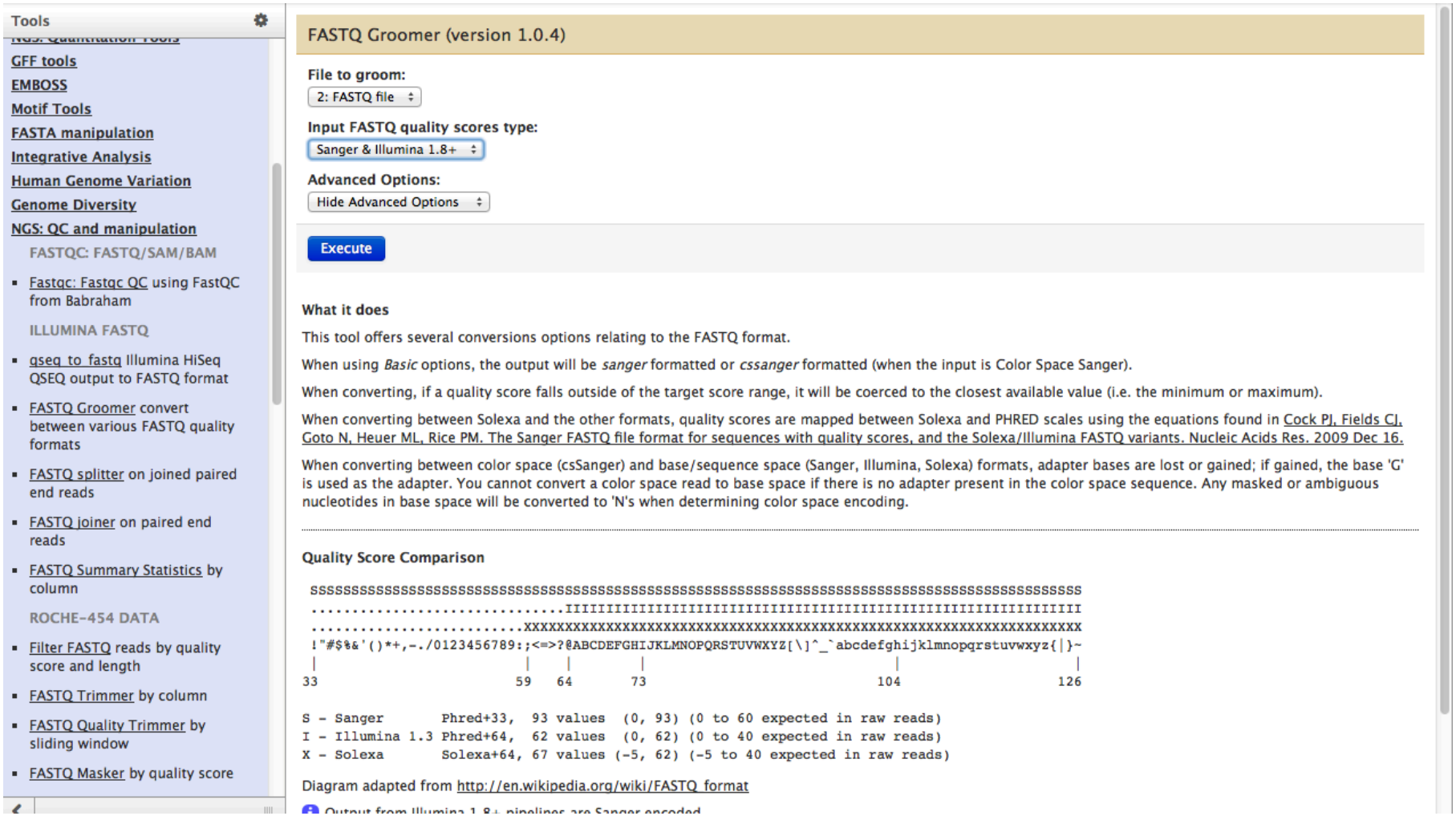
- ✓ Upload s_1_1_600000_qseq.txt
- ✓ Run qseq_to_fastq program
- ✓ Run Fastqc program



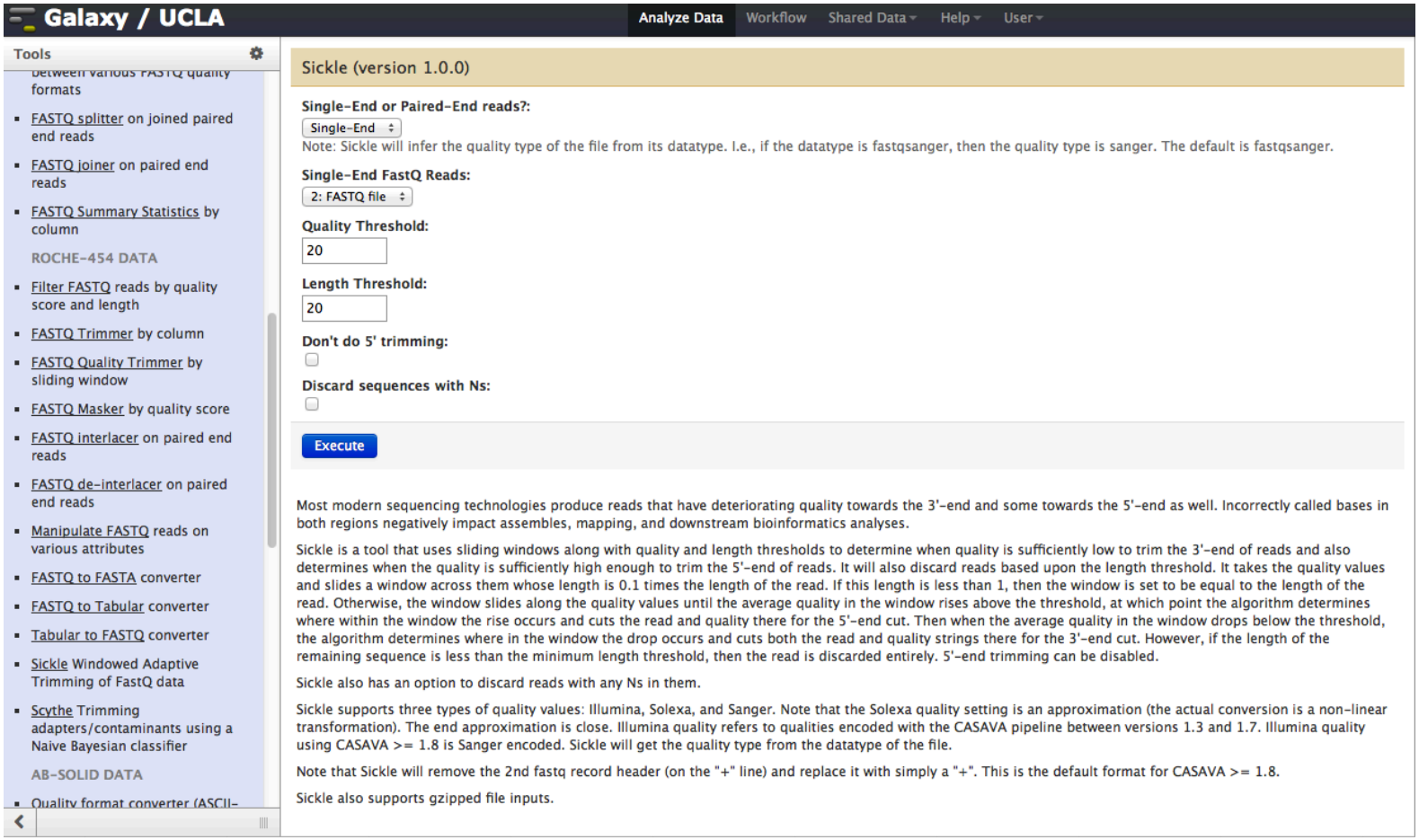


Alternatively, use compute quality statistics -> draw quality score boxplot -> draw nuclotides distribution chart programs

FastQ Converter

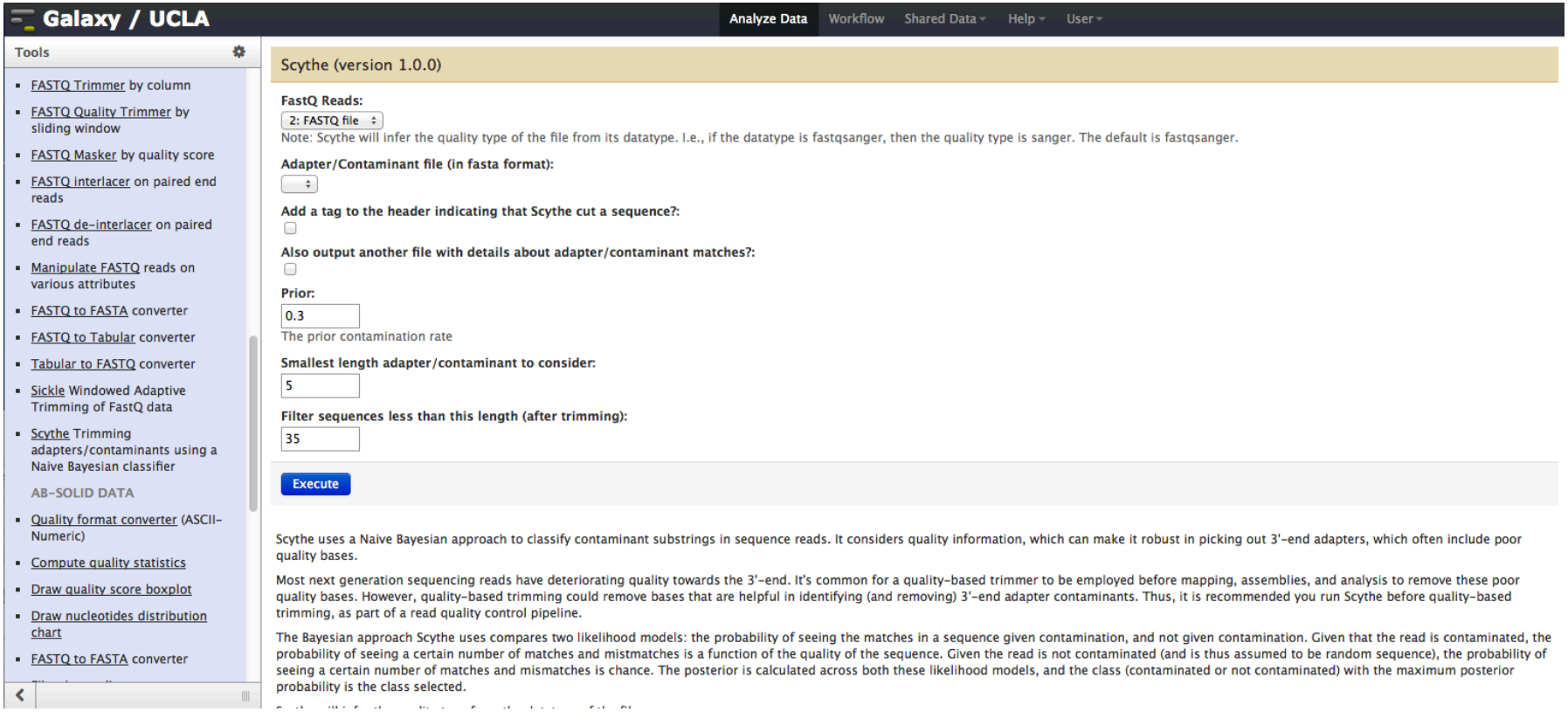


FastQ Manipulation



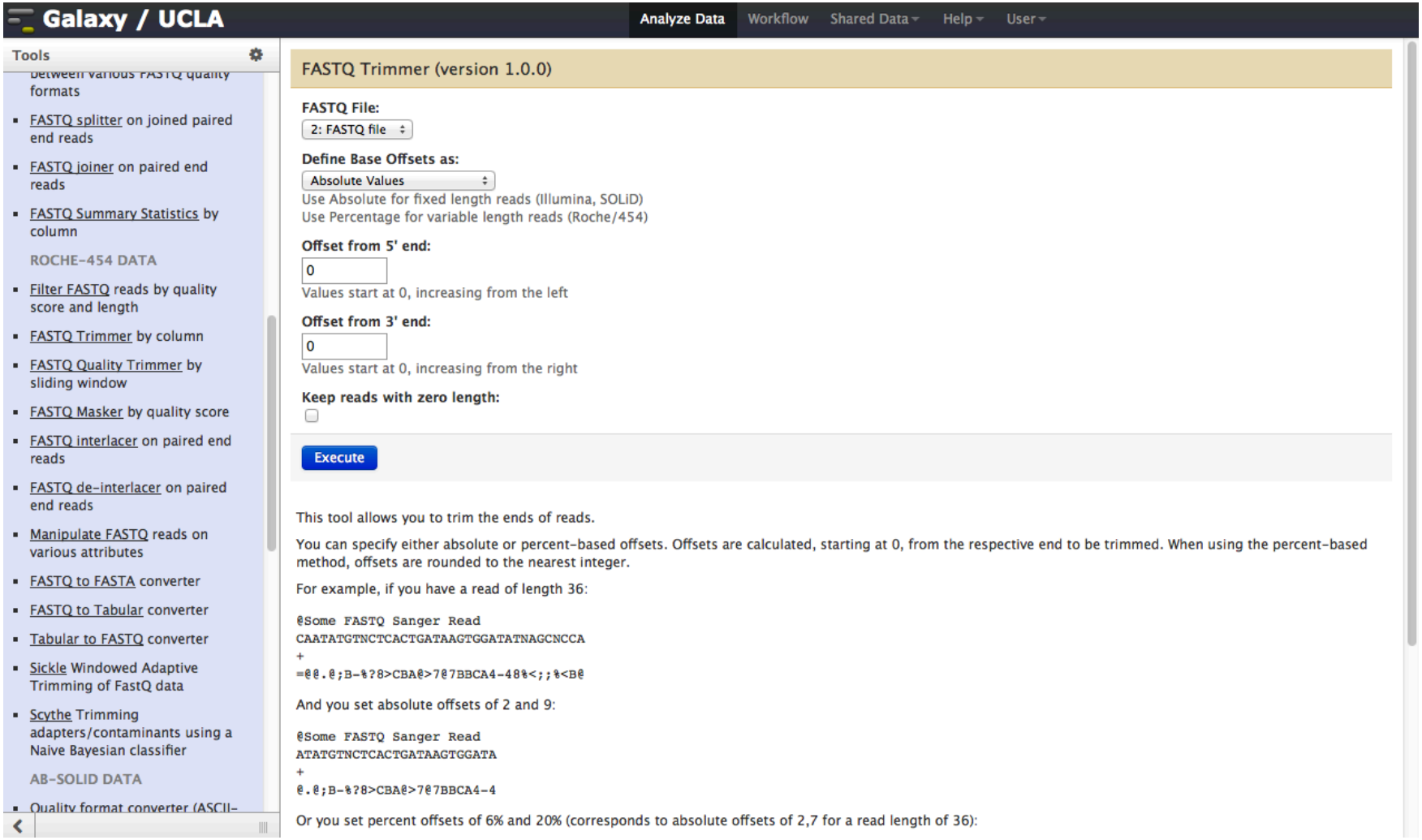
Sickle is a sliding window trimmer and tries to keep the longest high quality 5’ sequence reads.
windows of N bases moving from 5’ to 3’ end are tested for average quality. In the first window that fails to meet >Q, bases are trimmed starting with the first base with quality < Q

FastQ Manipulation



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

FastQ Manipulation



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 Insitute. 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.

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	CIAGR	MRNM	MPOS	ISIZE	SEQ	QUAL	OPT		
HWI-EAS91_1_30788AAXX:1:1:1761:343	4	*	0	0	*	*	0	0	AAAAAAAN				
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]
```

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

RNA-Seq Aligners

- ✓ Bowtie

  It doesn't perform gapped alignments. It runs faster and requires smaller memory footprint.

- ✓ Bowtie2

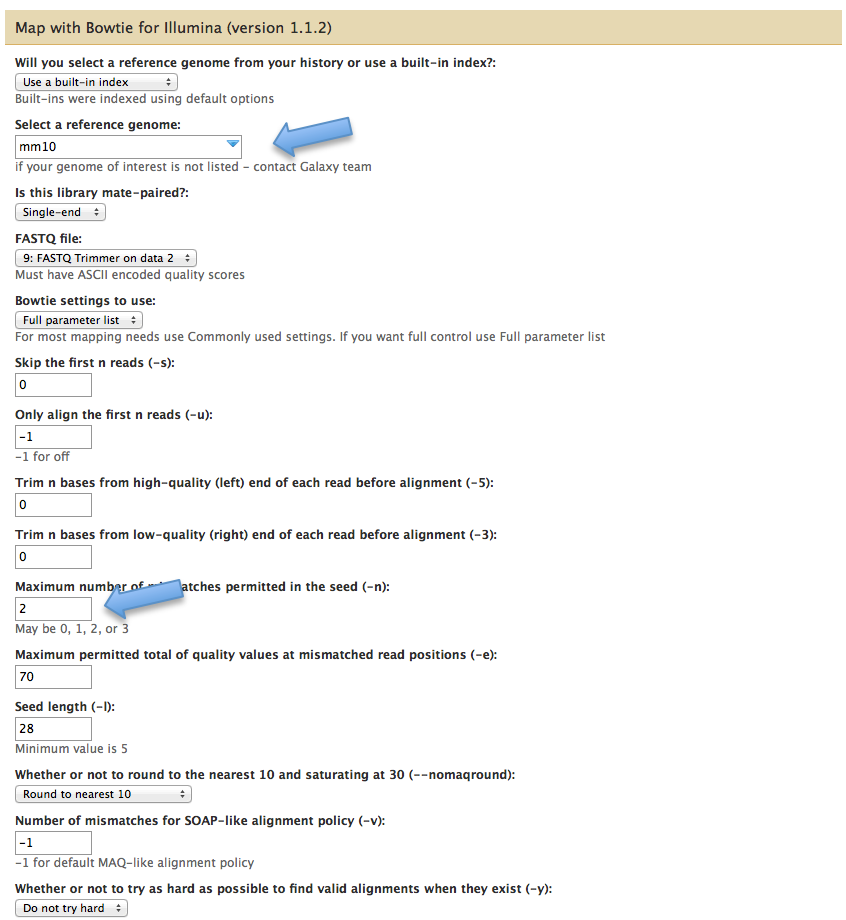
  It is fast and can perform local and gapped alignment. It performs better for reads longer than 50bp.

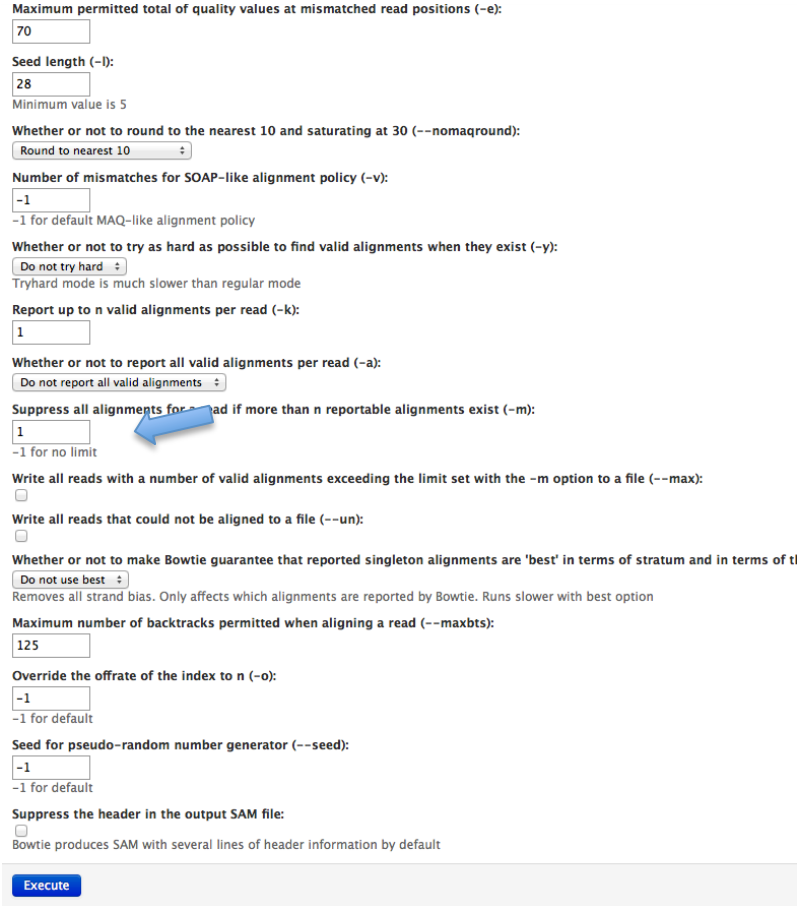
  Bowtie and bowtie2 use indexed reference genome

- ✓ 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





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.

```
Coor     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:

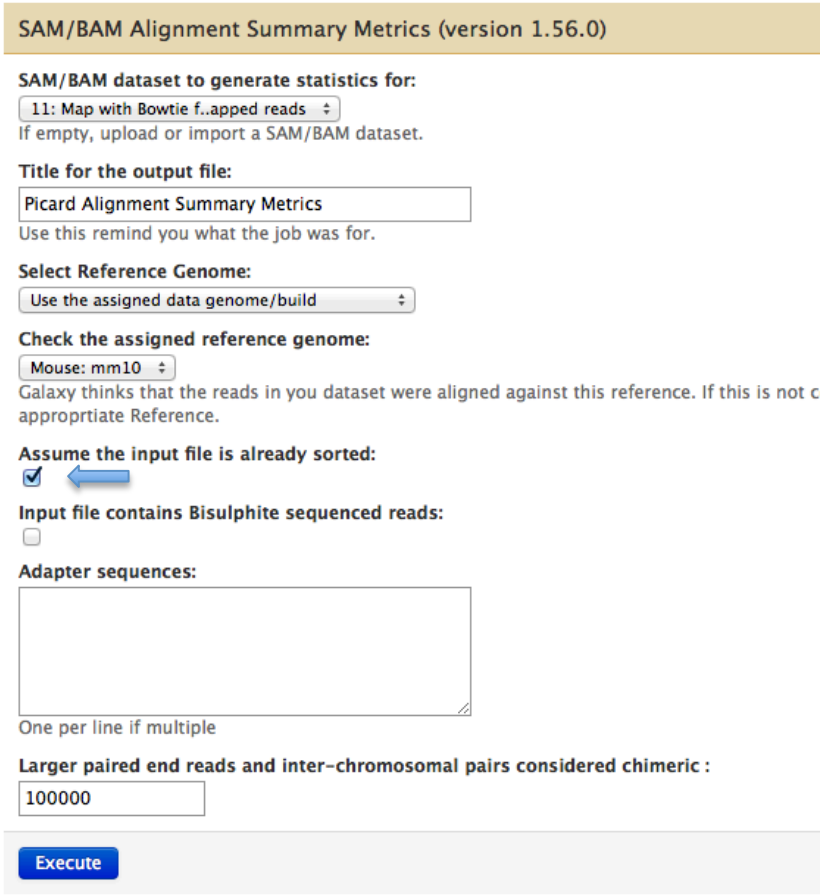
```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001   99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002    0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003    0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004    0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC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

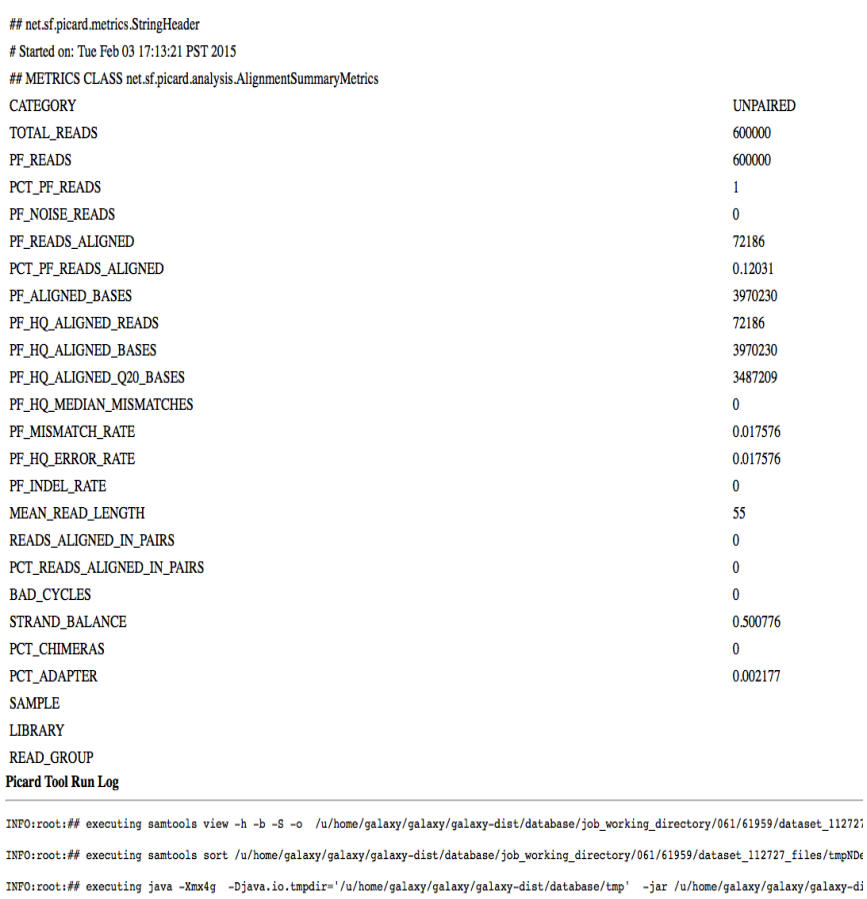
http://samtools.github.io/hts-specs/SAMv1.pdf

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





net.sf.picard.metrics.StringHeader
Started on: Tue Feb 03 17:13:21 PST 2015
METRICS CLASS net.sf.picard.analysis.AlignmentSummaryMetrics

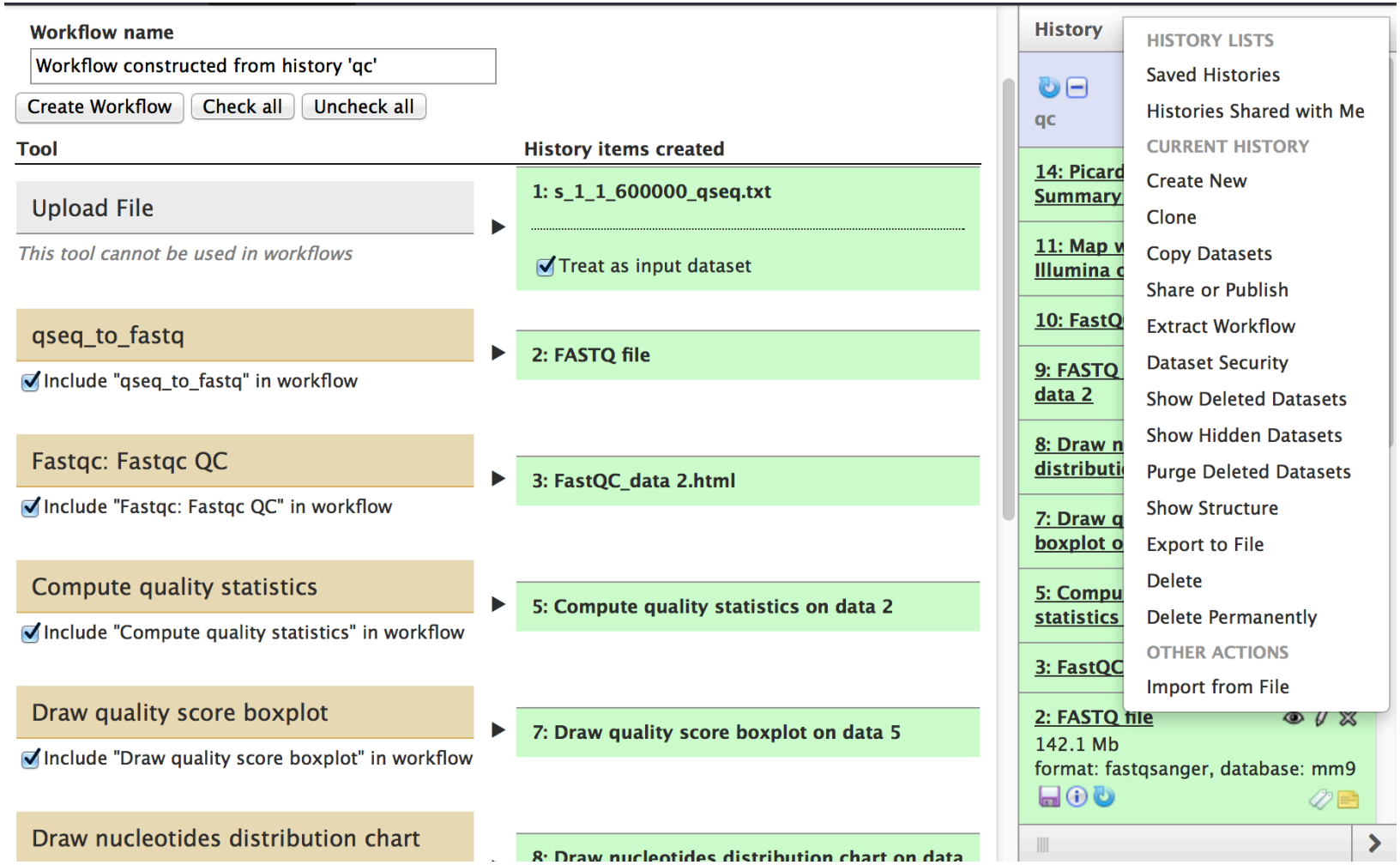
CATEGORY	UNPAIRED
TOTAL_READS	600000
PF_READS	600000
PCT_PF_READS	1
PF_NOISE_READS	0
PF_READS_ALIGNED	72186
PCT_PF_READS_ALIGNED	0.12031
PF_ALIGNED_BASES	3970230
PF_HQ_ALIGNED_READS	72186
PF_HQ_ALIGNED_BASES	3970230
PF_HQ_ALIGNED_Q20_BASES	3487209
PF_HQ_MEDIAN_MISMATCHES	0
PF_MISMATCH_RATE	0.017576
PF_HQ_ERROR_RATE	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

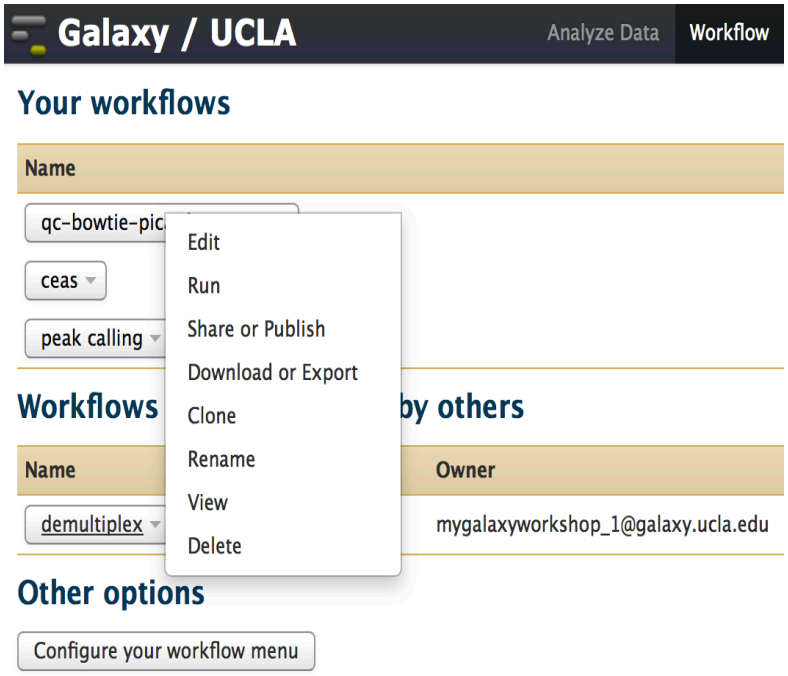
```
INFO:root:## executing samtools view -h -b -S -o  /u/home/galaxy/galaxy/galaxy-dist/database/job_working_directory/061/61959/dataset_112727
INFO:root:## executing samtools sort /u/home/galaxy/galaxy/galaxy-dist/database/job_working_directory/061/61959/dataset_112727_files/tmpNDe
INFO:root:## executing java -Xmx4g  -Djava.io.tmpdir='/u/home/galaxy/galaxy/galaxy-dist/database/tmp'  -jar /u/home/galaxy/galaxy/galaxy-di
```

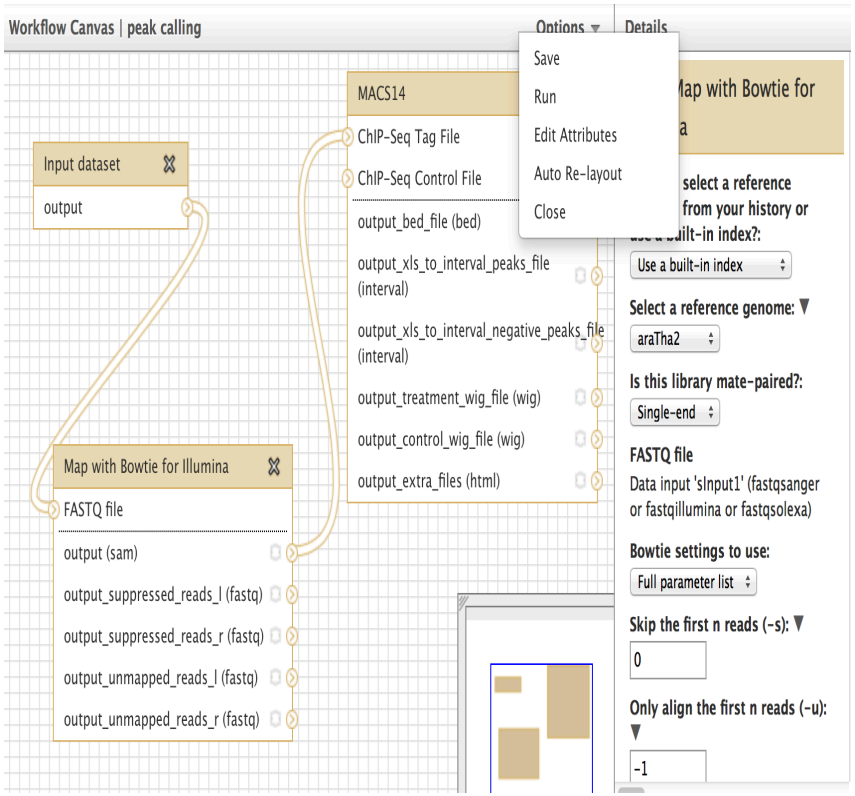
Uncheck “assume the input file is already sorted”

Extract Workflow



Workflow Management

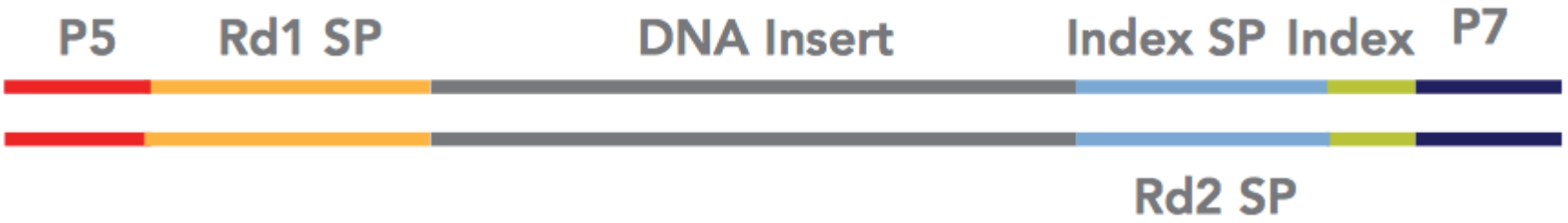




Published workflow/history listed as shared data

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

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.

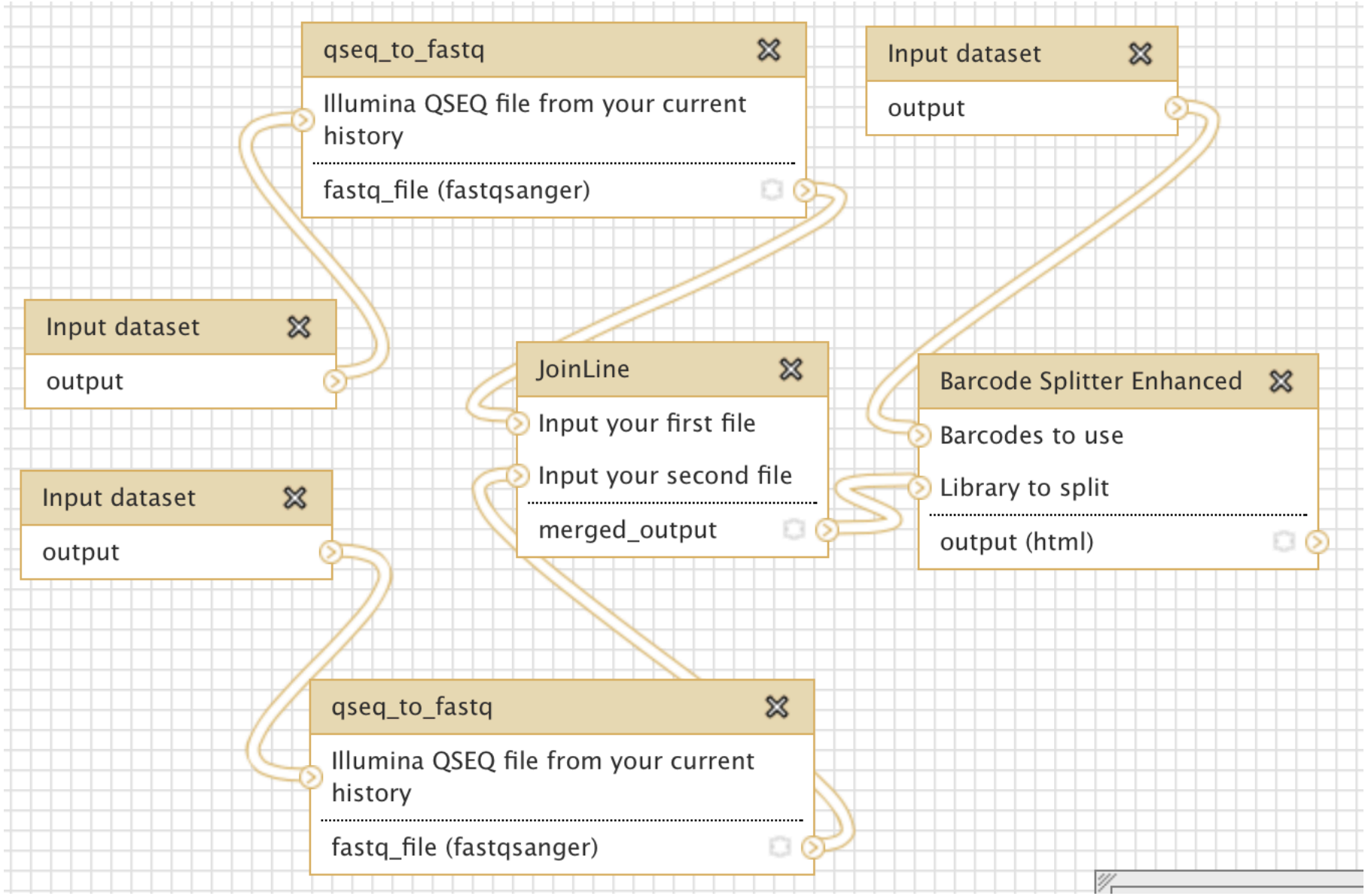
- ✓ Demultiplex workflow

  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
CAGCTCATTGATGCAGTCCAGGCACTTCCCCCACATCTCTTCATGTAGGT
+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
GATTAGGGTGCTTAGCTGTTAACTAAGCGTTTGTGGGTTTAAGTCCCATG
+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.