Introduction to HPRC Galaxy

Your Login Password

- Both state of Texas law and TAMU regulations prohibit the sharing and/or illegal use of computer passwords and accounts
- Don’t write down passwords
- Don’t choose easy to guess/crack passwords
- Change passwords frequently

For More Help...

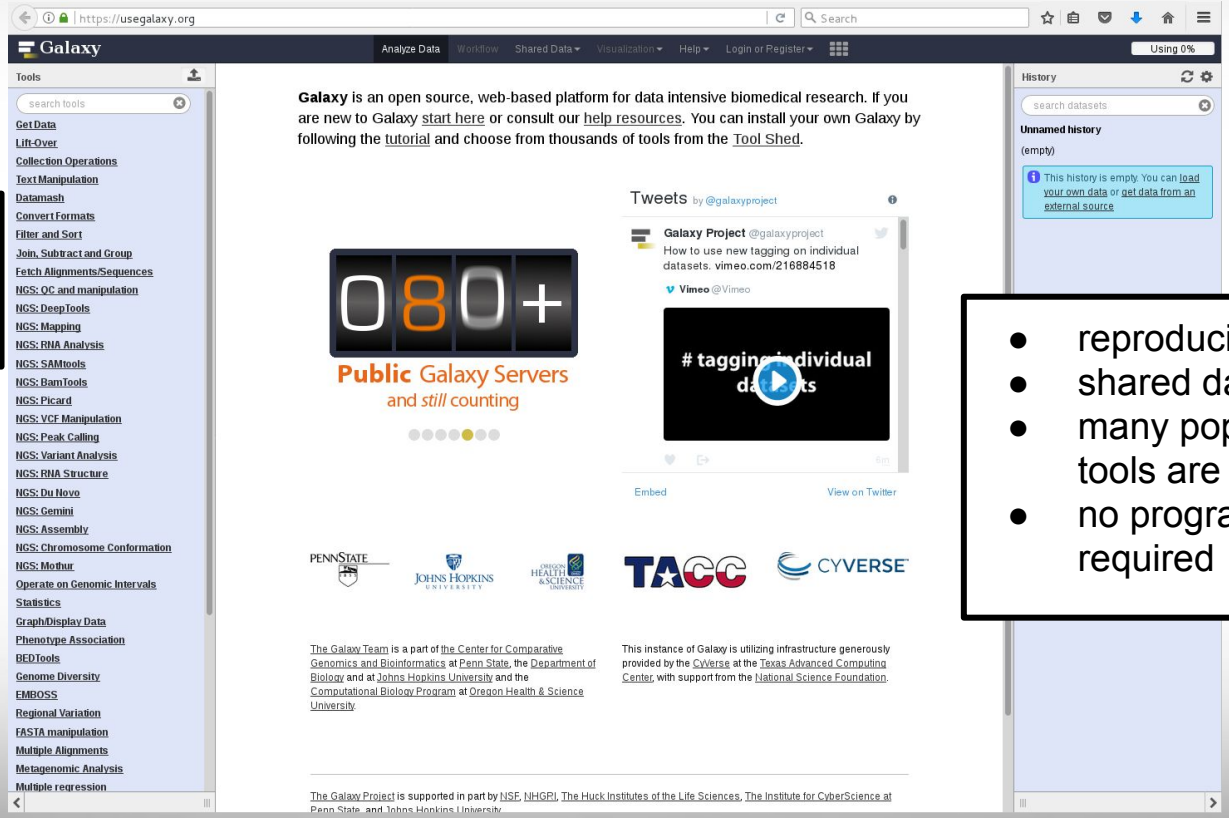
Website:	hprc.tamu.edu
Email:	help@hprc.tamu.edu
Telephone:	(979) 845-0219
Visit us in person:	Henderson Hall, Room 114A

Help us, help you -- we need more info

- Which Cluster
- UserID/NetID
- Job id(s) if any
- Location of your jobfile, input/output files
- Application used if any
- Module(s) loaded if any
- Error messages
- Steps you have taken, so we can reproduce the problem

What is the Galaxy Project?

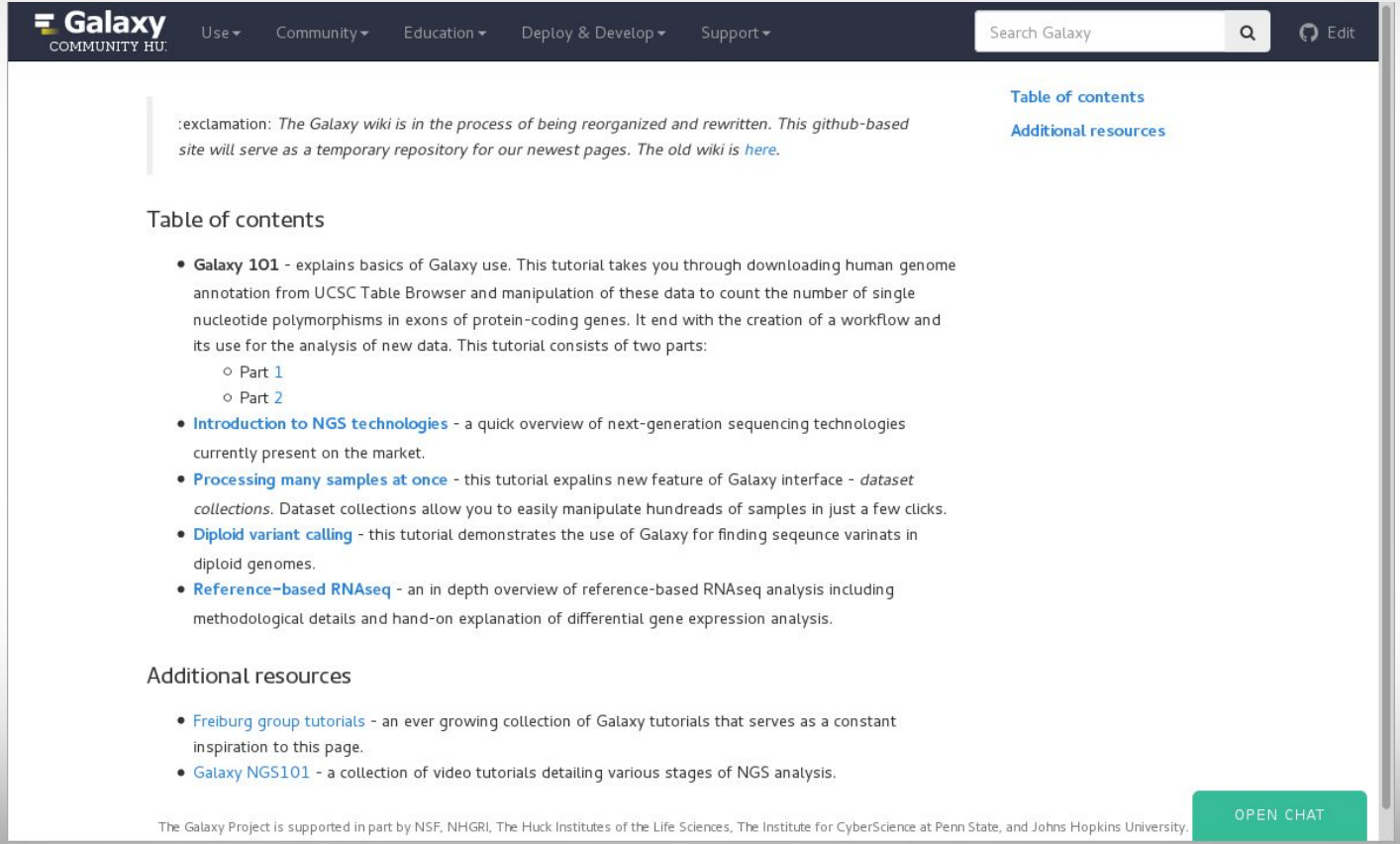
HPRC Galaxy is not a public Galaxy server



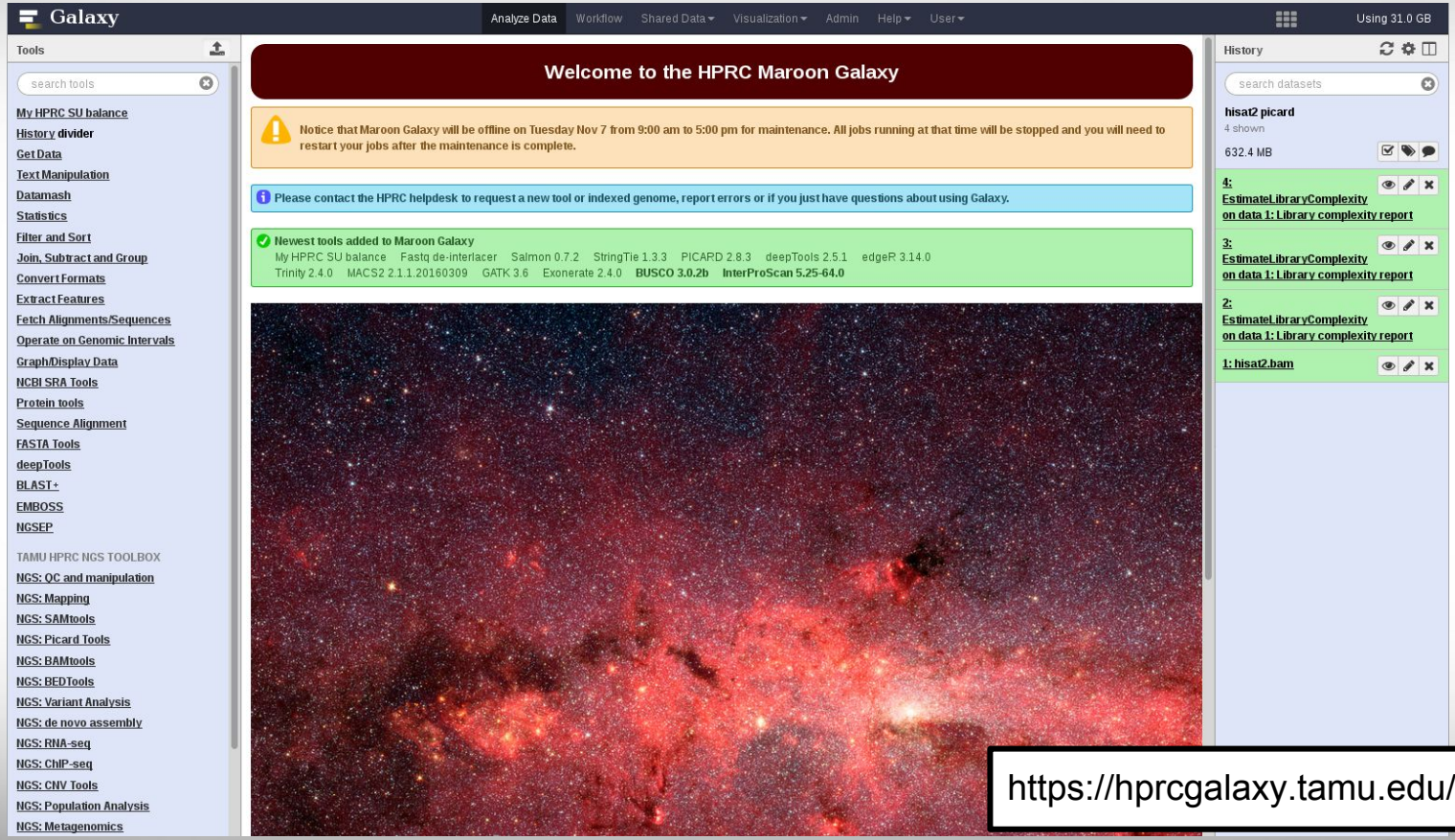
usegalaxy.org

- reproducible workflow
- shared data and workflows
- many popular bioinformatic tools are available
- no programming knowledge required

Galaxy 101 galaxyproject.org/tutorials



HPRC Maroon Galaxy



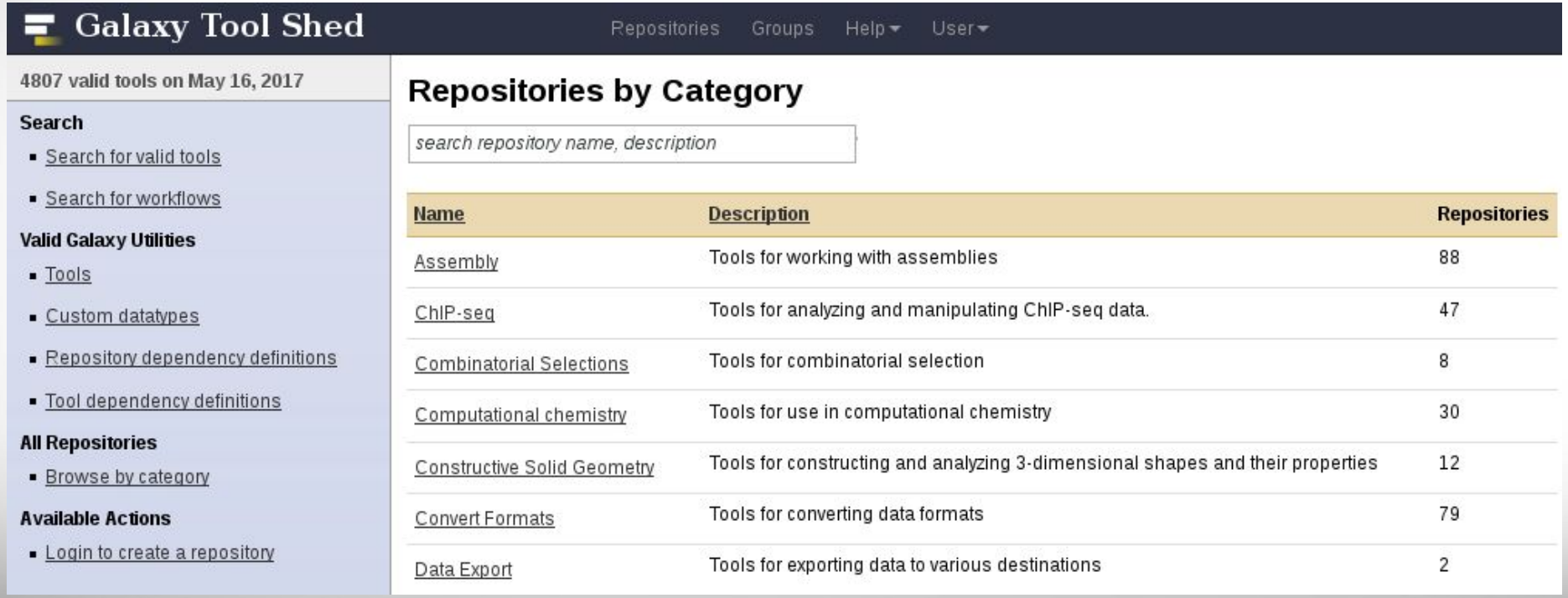
https://hprcgalaxy.tamu.edu/maroon/

HPRC Maroon Galaxy

- Try Galaxy at usegalaxy.org to see if it appropriate for your project
- Getting Access to HPRC Maroon Galaxy
  - Available to Texas A&M students, staff and faculty with a NetID and Ada account
  - Apply for an Ada account first
    - https://hprc.tamu.edu/apply
  - Then send an email request for a Maroon Galaxy account
    - help@hprc.tamu.edu
- Read the Galaxy Usage Notes
  - https://hprc.tamu.edu/wiki/Ada:Galaxy
- There are no backups of users' Galaxy files
  - You can export a Galaxy history to a single file which can be uploaded to the same or a different Galaxy instance
    - Some custom tools may not be available in all Galaxies

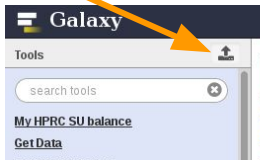
Look for Additionally Available Tools

toolshed.g2.bx.psu.edu



Get Data into Your Galaxy History

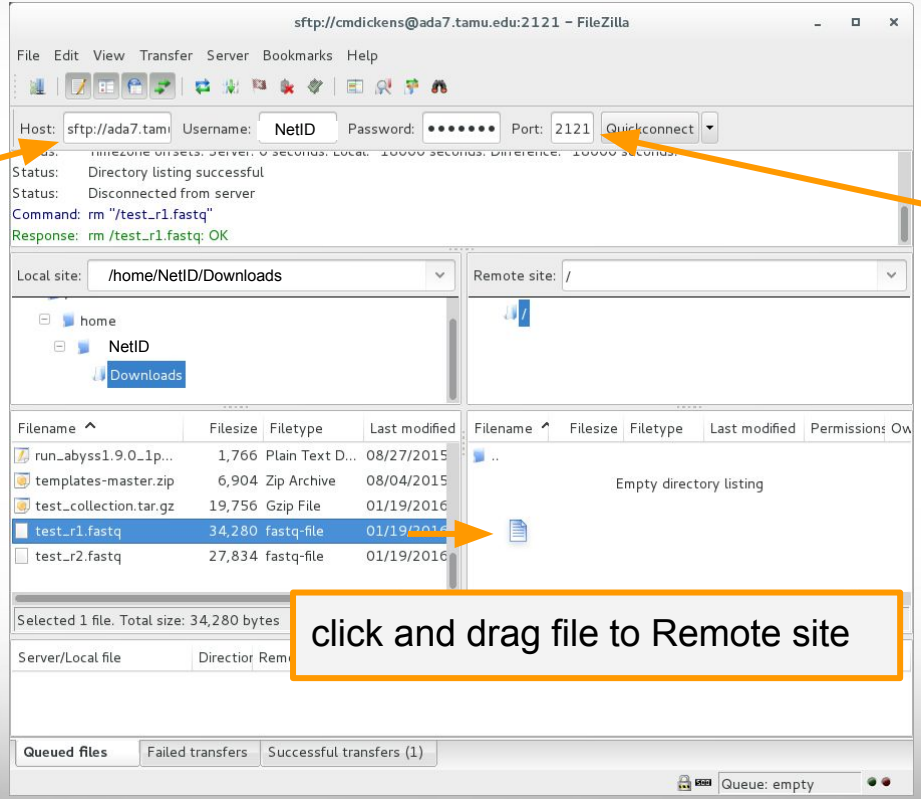
- Upload files < 2GB using Galaxy web interface
  - select local file to upload or paste URL
- Upload large files > 2GB
  - using sftp UNIX utility
  - using sftp with locally installed Filezilla or WinSCP
  - paste URL if it is a FTP URL
- Can retrieve data from external websites directly into your Galaxy history with ‘Get Data’ tools
  - UCSC, BioMart, Ratmine, ...



Copy a 2GB+ desktop file to Galaxy with Filezilla

Host: sftp://ada7.tamu.edu

for file sizes > 2GB

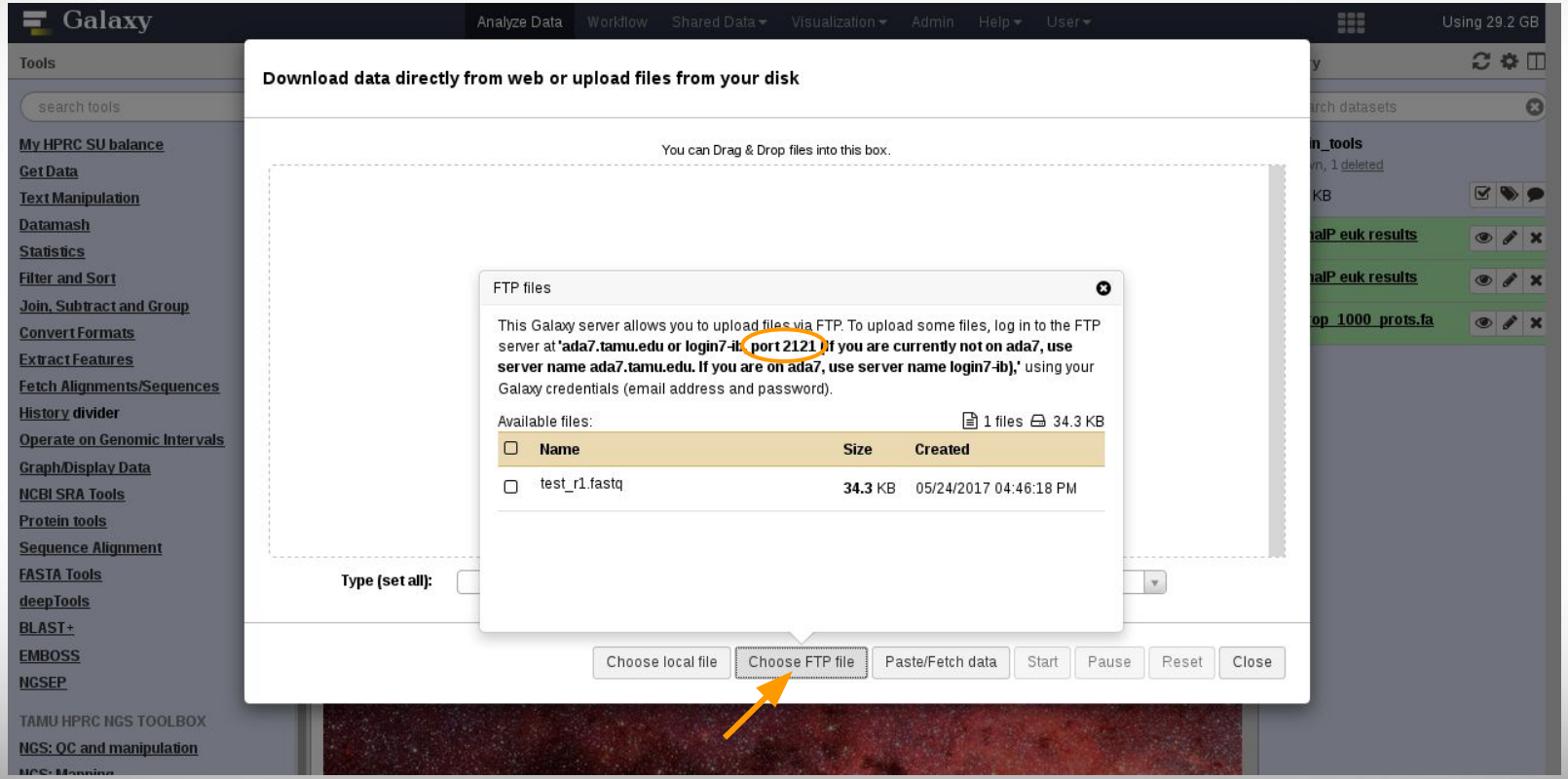


Port: 2121

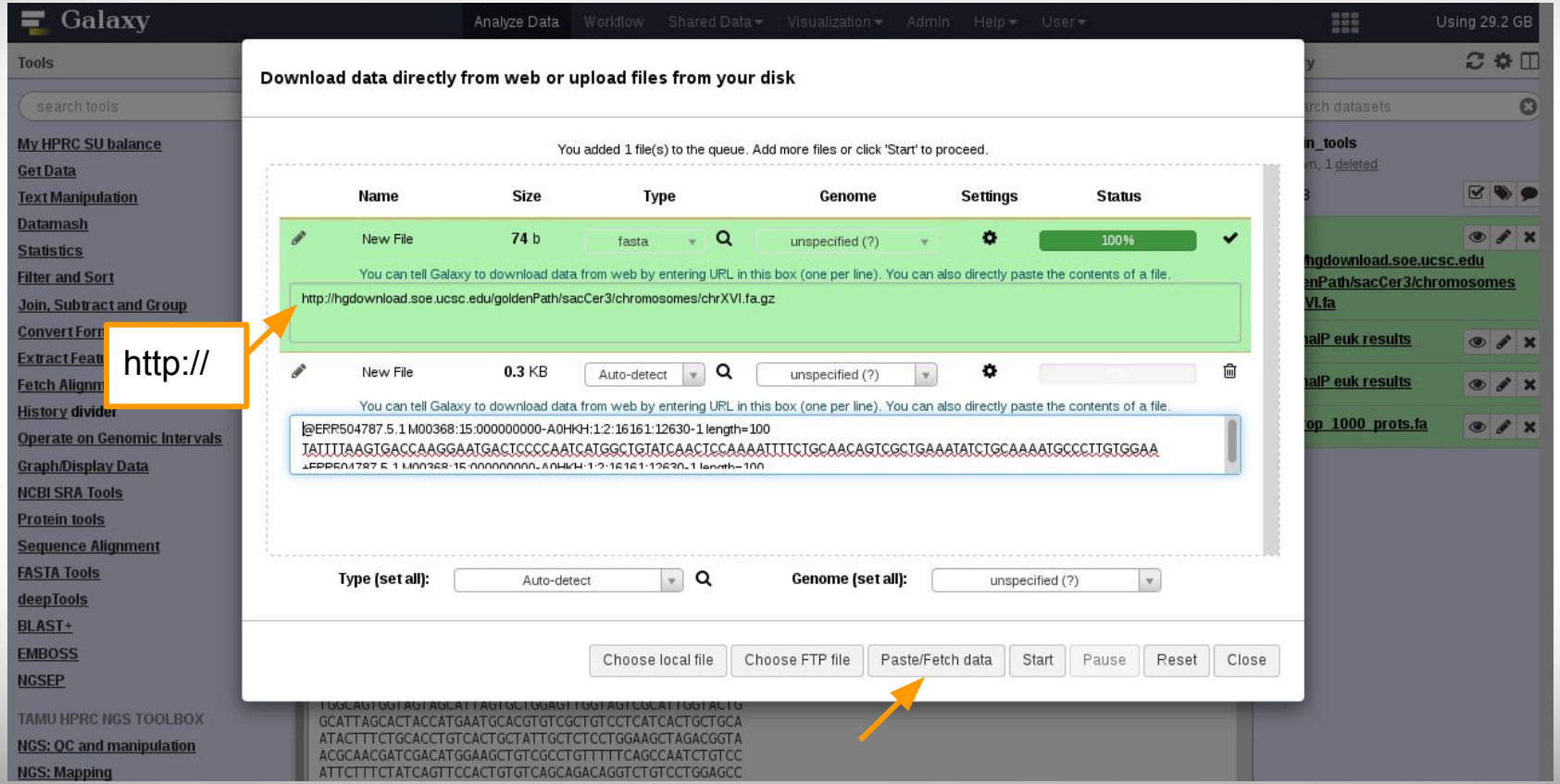
click and drag file to Remote site

After transferring file to the 'Remote site' directory, go to Galaxy upload file interface to see your ftp transferred file (next slide)

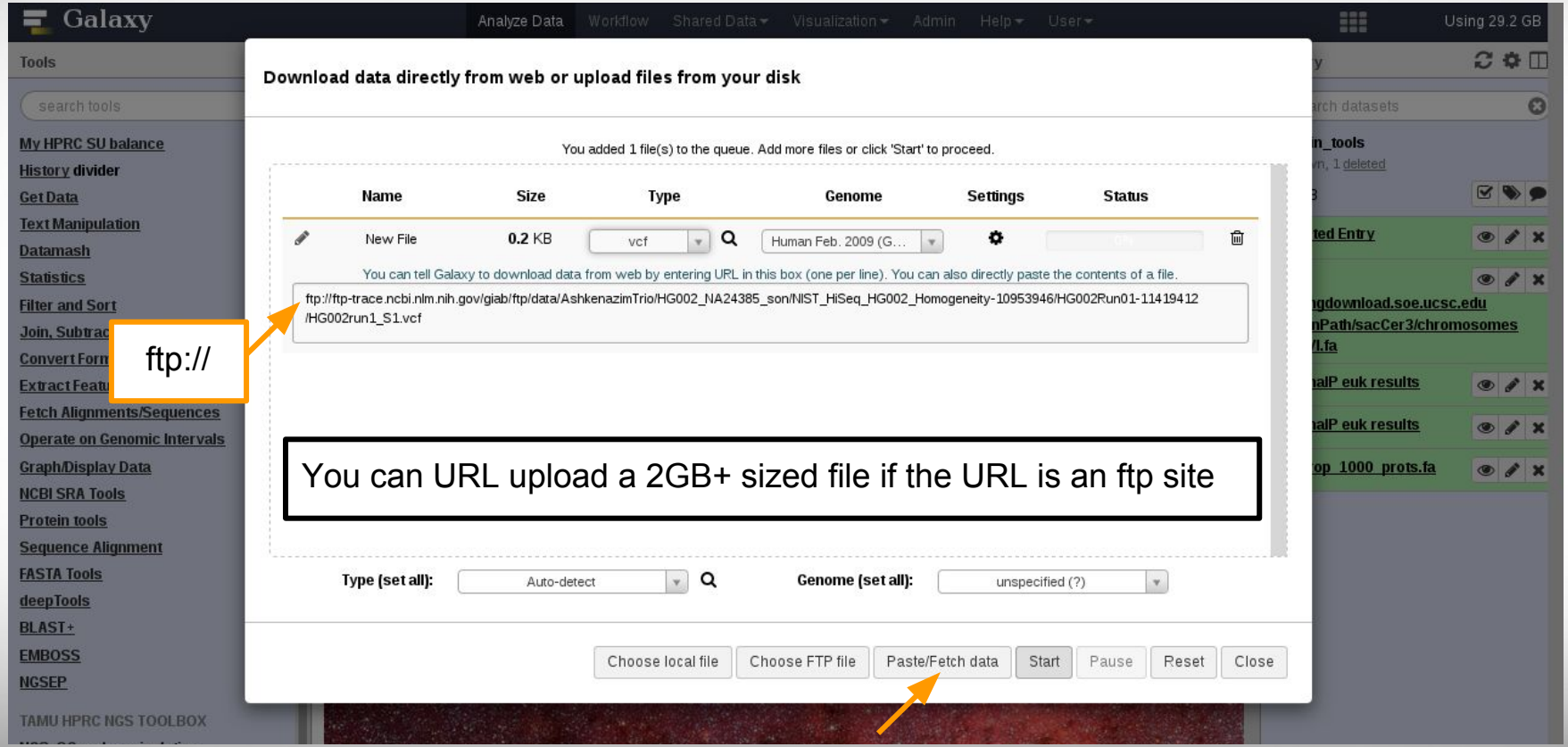
Add Your FTP Uploaded 2GB+ File to Your History



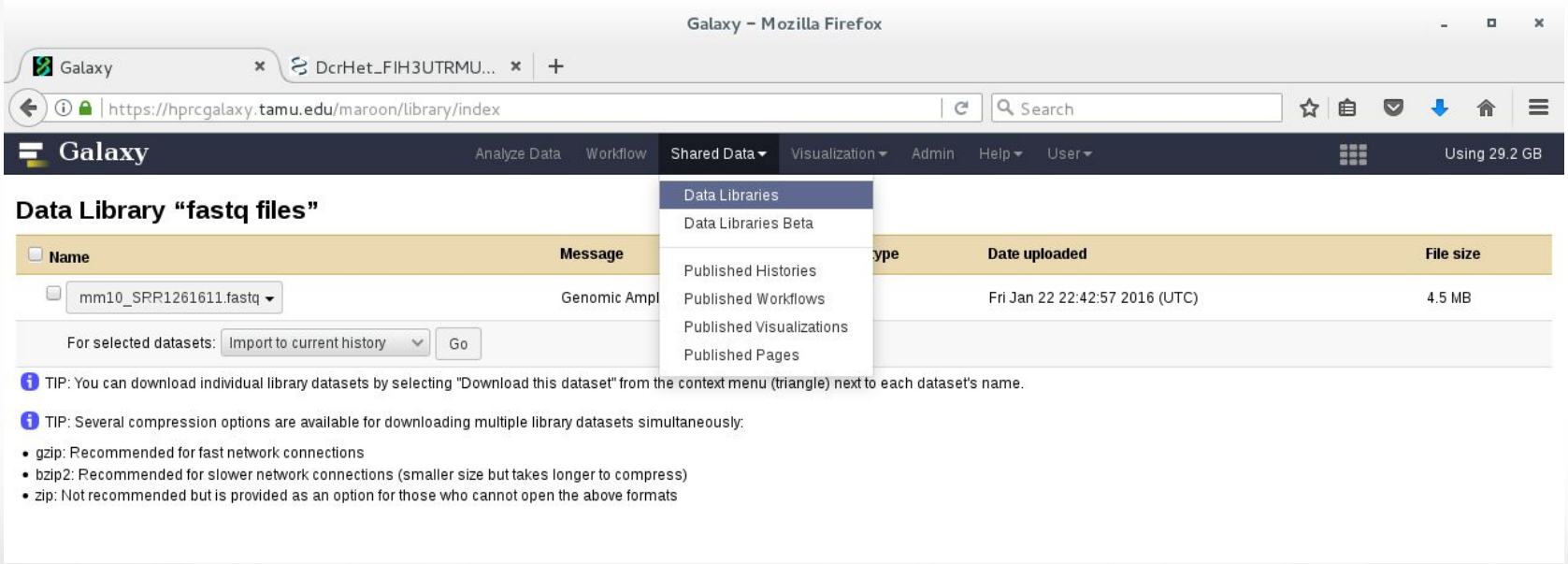
HTTP URL Upload < 2GB File or Direct Paste



FTP URL Upload 2GB+ File via URL



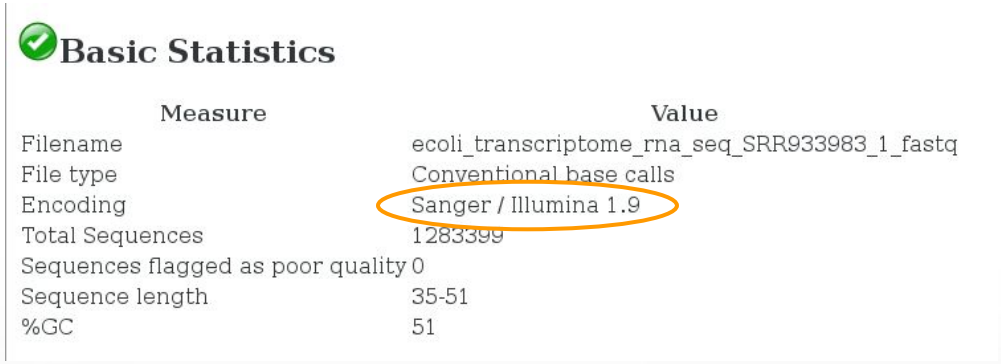
Shared Data Libraries



Files can be added to a 'Data Library' which you can share with your colleagues. Send a request to help@hprc.tamu if you would like to create a Data Library

Galaxy File Formats

- Many Galaxy tools require fastqsanger format
- FastQC tool will check fastq format of a fastq file
  - MiSeq, HiSeq, NextSeq, NovaSeq all use 1.8+

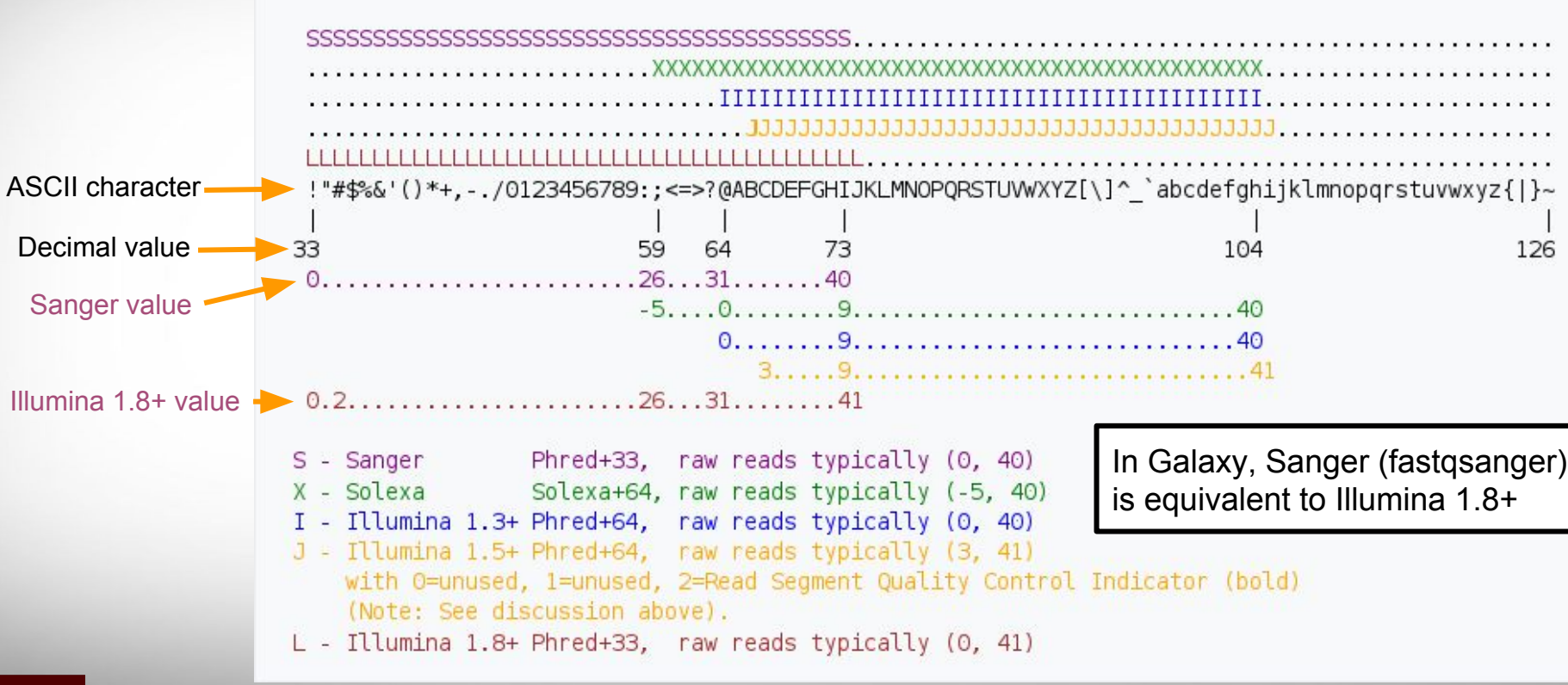


Basic Statistics

Measure	Value
Filename	ecoli_transcriptome_rna_seq_SRR933983_1_fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1283399
Sequences flagged as poor quality	0
Sequence length	35-51
%GC	51

```
@ERR504787.5.1 M00368:15:000000000-A0HKH:1:2:16161:12630-1 length=100
TATTTTAAGTGACCAAGGAATGACTCCCCAATCATGGCTGTATCAACTCCAAAATTTTCTGCAACAGTCGCTGAAATATCTGCAAAATGCCCTTGTGGAA
+ERR504787.5.1 M00368:15:000000000-A0HKH:1:2:16161:12630-1 length=100
CCCFFFFFHHHHHJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJIJJJHIJHHGFFFFFFFEEEEEEDDDDDDDDDCDDDD
```

FASTQ Format Encoding

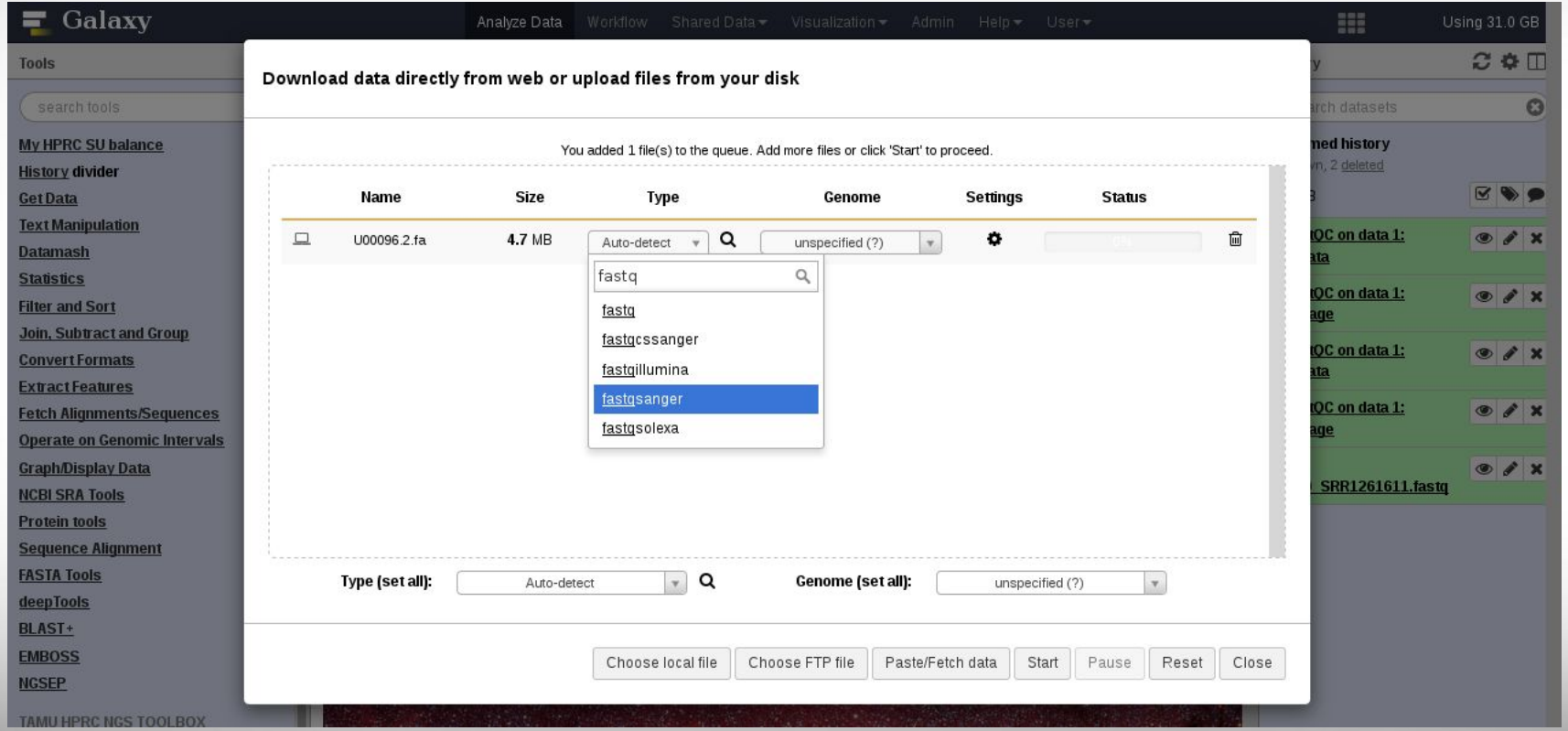


In Galaxy, Sanger (fastqsanger) is equivalent to Illumina 1.8+

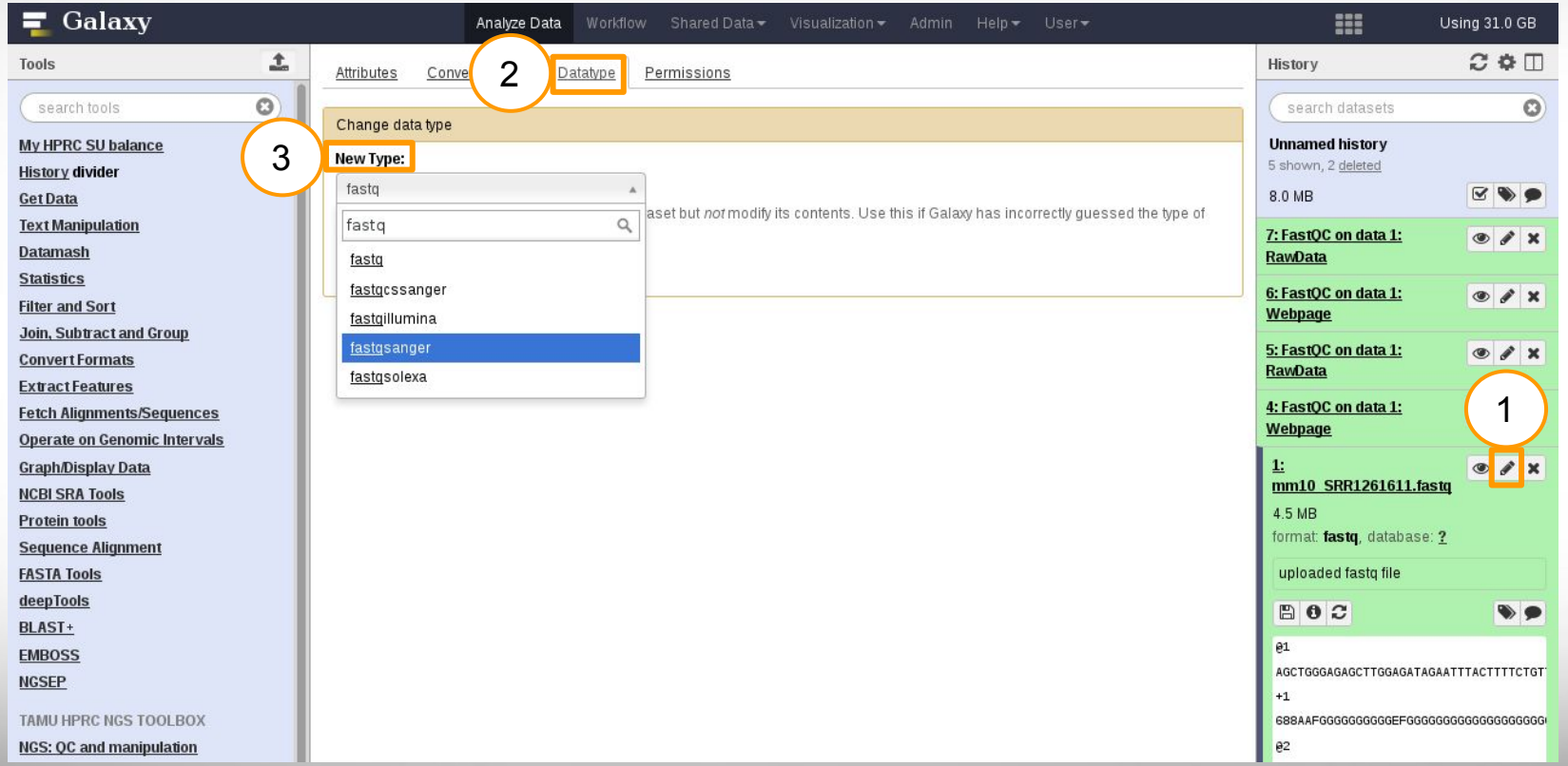
Galaxy File Formats

- Fastq Groomer tool will convert fastq to fastqsanger
  - Will convert quality scores to fastqsanger encoding
    - Solexa to fastqsanger
    - Illumina 1.3 - 1.7 to fastqsanger
  - converted file will be an exact copy if the fastq file is already Illumina 1.8+
- If your fastq file is already in fastqsanger format, you can set format in file attributes instead of making a copy which will save you time and also save disk space

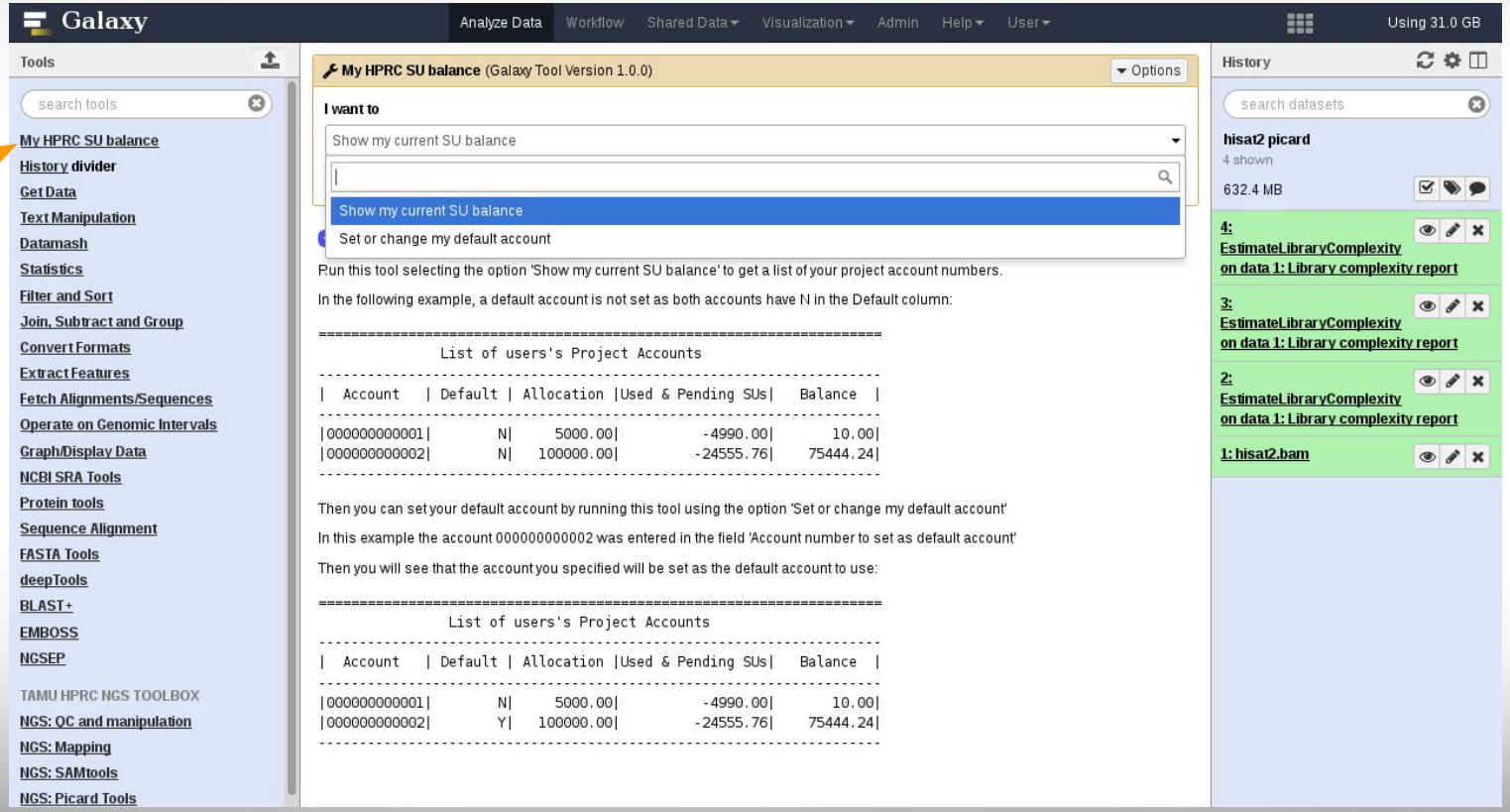
Setting Format to Fastqsanger on Upload



Setting Format to Fastqsanger in File Attributes



Check Your HPRC SUs Balance



BLAST+ Multiple Walltime Tools

====== 1 DAY JOBS ======

NCBI BLAST+ blastn 480 SUs. Search nucleotide database with nucleotide query sequence(s) (max runtime 1 day, 480 SUs required)

NCBI BLAST+ blastp 480 SUs. Search protein database with protein query sequence(s) (max runtime 1 day, 480 SUs required)

NCBI BLAST+ blastx 480 SUs. Search protein database with translated nucleotide query sequence(s) (max runtime 1 days, 480 SUs required)

NCBI BLAST+ tblastn 480 SUs Search translated nucleotide database with protein query sequence(s) (max runtime 1 days, 480 SUs required)

NCBI BLAST+ tblastx 480 SUs. Search translated nucleotide database with translated nucleotide query sequence(s) (max runtime 1 day, 480 SUs required)

BLAST Reciprocal B
from two FASTA files
1 day, 480 SUs req

====== 3 DAY JOBS ======

NCBI BLAST+ blastn 1440 SUs. Search nucleotide database with nucleotide query sequence(s) (max runtime 3 days, 1440 SUs required)

NCBI BLAST+ blastp 1440 SUs.Search protein database with protein query sequence(s) (max runtime 3 days, 1440 SUs required)

NCBI BLAST+ blastx 1440 SUs. Search protein database with translated nucleotide query sequence(s) (max runtime 3 days, 1440 SUs required)

NCBI BLAST+ tblastn 1440 SUs. Search translated nucleotide database with protein query sequence(s) (max runtime 3 days, 1440 SUs required)

NCBI BLAST+ tblastx 1440 SUs. Search translated nucleotide database with translated nucleotide query sequence(s) (max runtime 3 days, 1440 SUs required)

====== 7 DAY JOBS ======

NCBI BLAST+ blastn 3360 SUs. Search nucleotide database with nucleotide query sequence(s) (max runtime 7 days, 3360 SUs required)

NCBI BLAST+ blastp 3360 SUs.Search protein database with protein query sequence(s) (max runtime 7 days, 3360 SUs required)

NCBI BLAST+ blastx 3360 SUs. Search protein database with translated nucleotide query sequence(s) (max runtime 7 days, 3360 SUs required)

NCBI BLAST+ tblastn 3360 SUs. Search translated nucleotide database with protein query sequence(s) (max runtime 7 days, 3360 SUs required)

NCBI BLAST+ tblastx 3360 SUs. Search translated nucleotide database with translated nucleotide query sequence(s) (max runtime 7 days, 3360 SUs required)

== 7 DAY 5 NODE JOBS ==

NCBI BLAST+ blastp tamulauncher 16,800 SUs. Search protein database with protein query sequence(s) (max runtime 7 days on 5 nodes, 16,800 SUs required)

NCBI BLAST+ blastn tamulauncher 16,800 SUs. Search nucleotide database with nucleotide query sequence(s) (max runtime 7 days, 5 nodes, 16,800 SUs required)

NCBI BLAST+ blastx 16,800 SUs. Search protein database with translated nucleotide query sequence(s) (max runtime 7 days on 5 nodes, 16,800 SUs required)

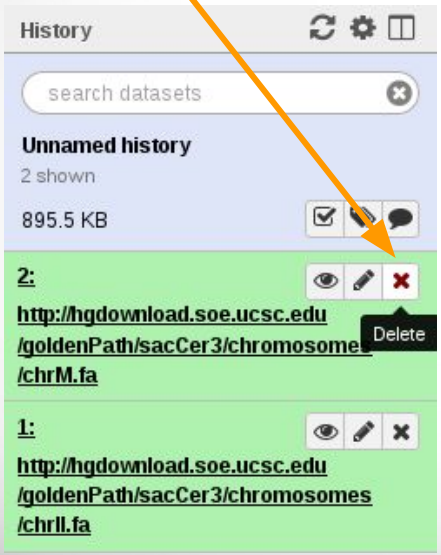
NCBI BLAST+ tblastn 16,800 SUs. Search translated nucleotide database with protein query sequence(s) (max runtime 7 days, 5 nodes, 16,800 SUs required)

NCBI BLAST+ tblastx 16,800 SUs. Search translated nucleotide database with translated
eotide query sequence(s)
runtime 7 days on 5 nodes,
00 SUs required)

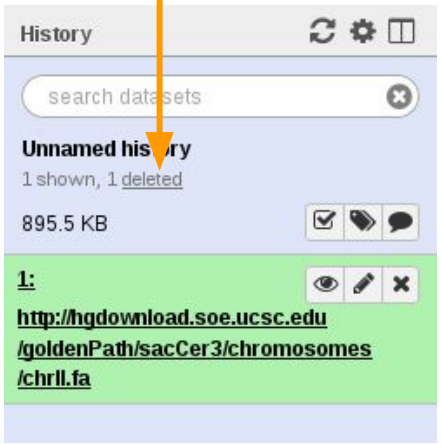
Galaxy jobs cannot be restarted and checkpoints are not supported

Permanently Delete Unwanted Files

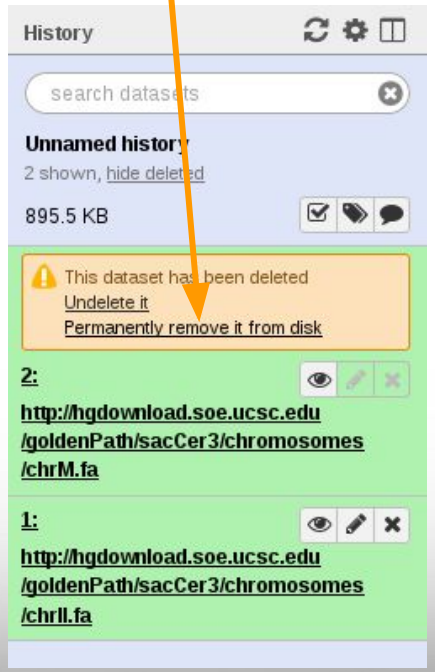
delete file



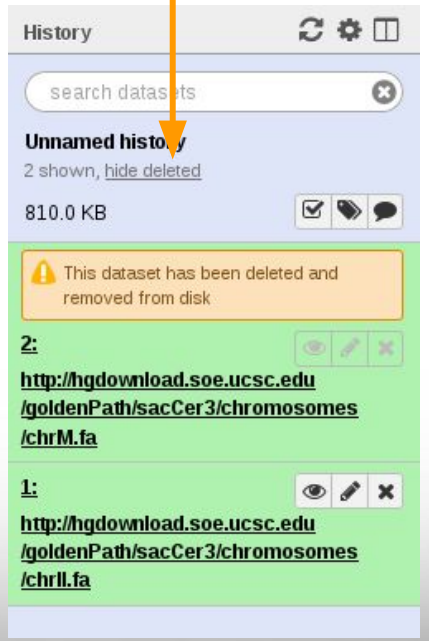
show deleted files



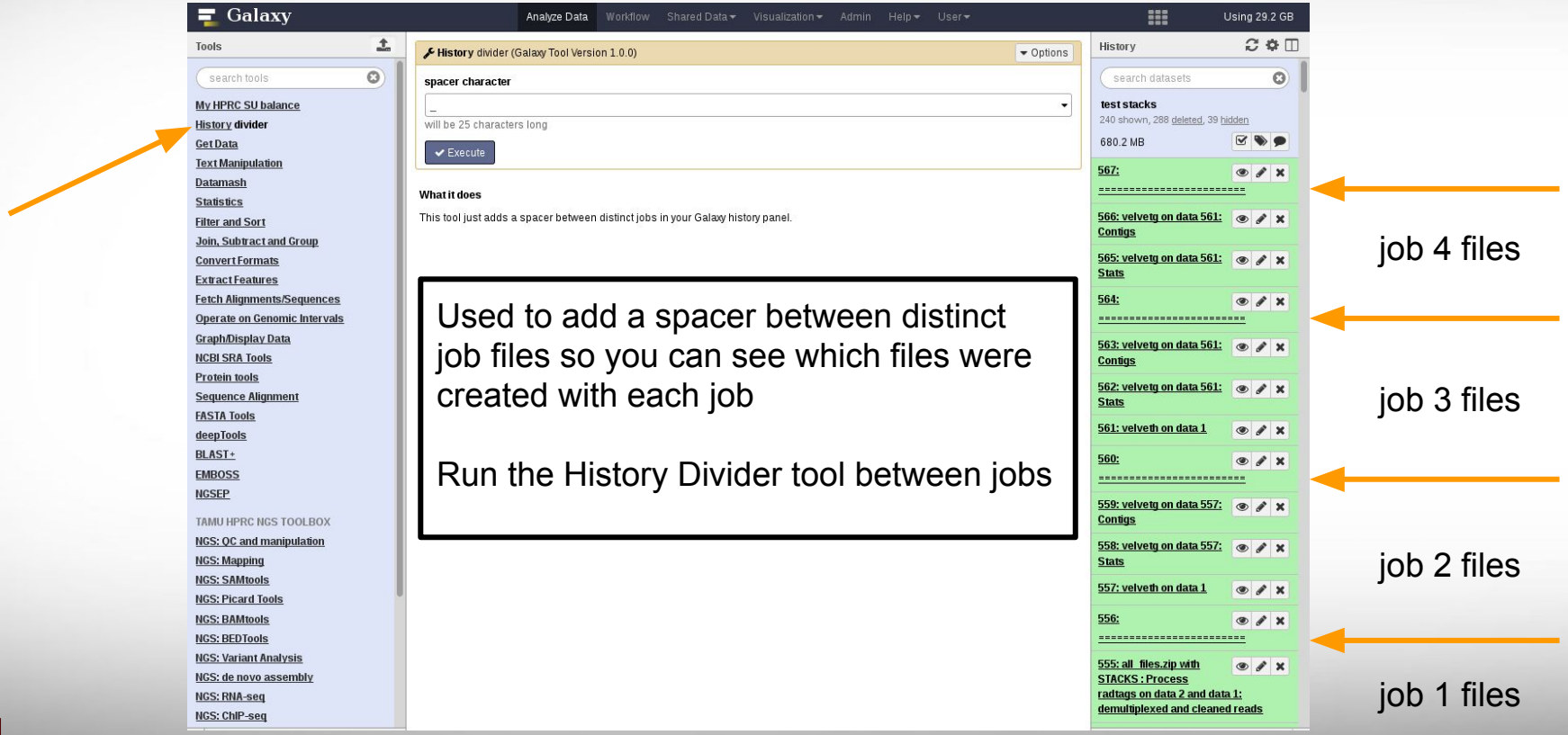
permanently delete file



hide deleted files



History Divider



Used to add a spacer between distinct job files so you can see which files were created with each job

Run the History Divider tool between jobs

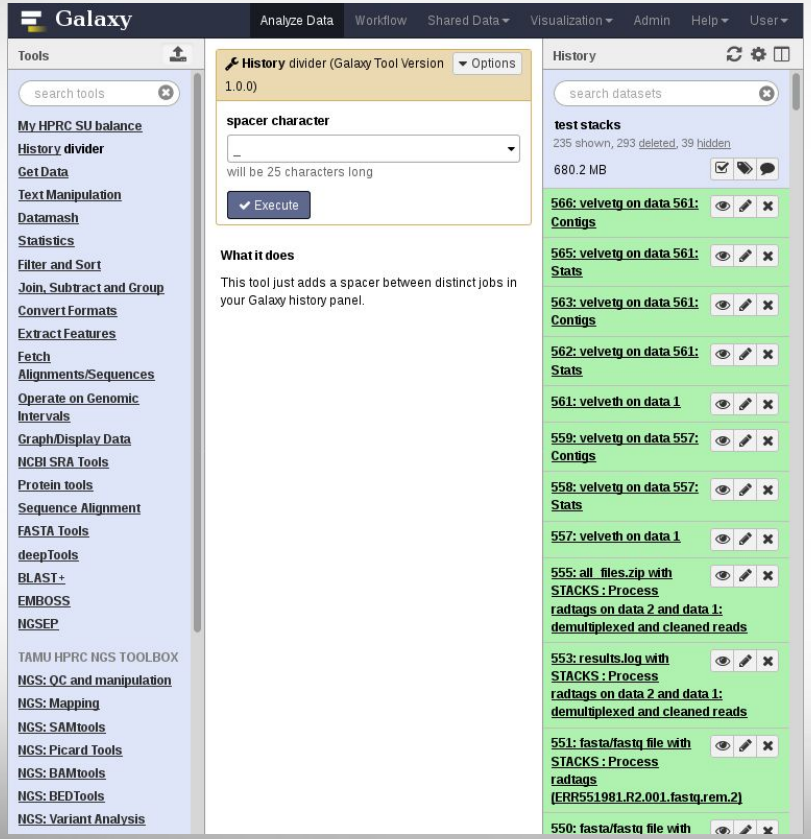
job 4 files

job 3 files

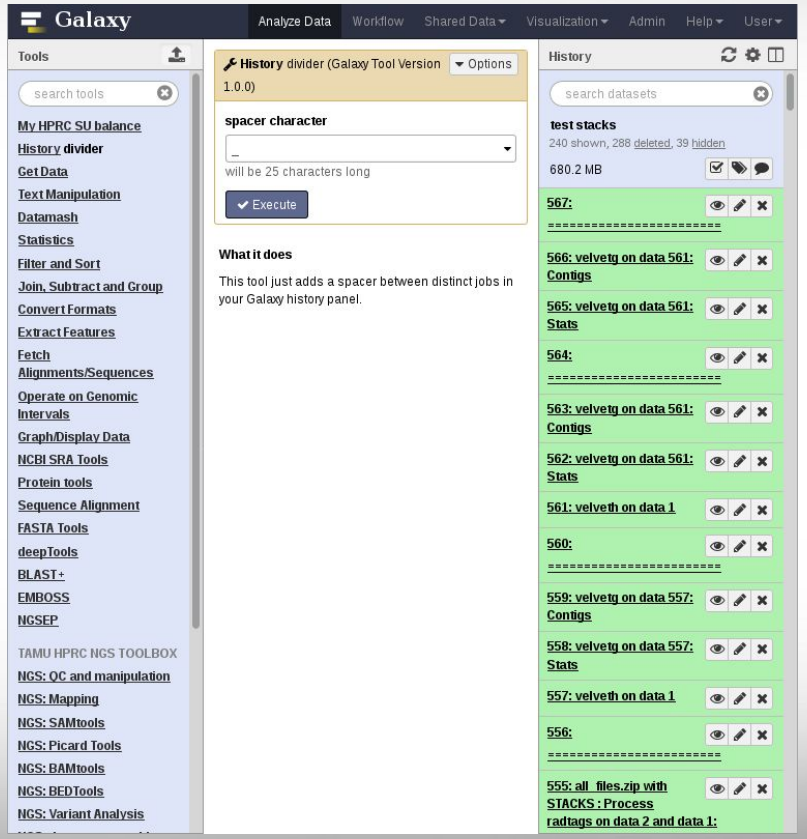
job 2 files

job 1 files

cluttered history items

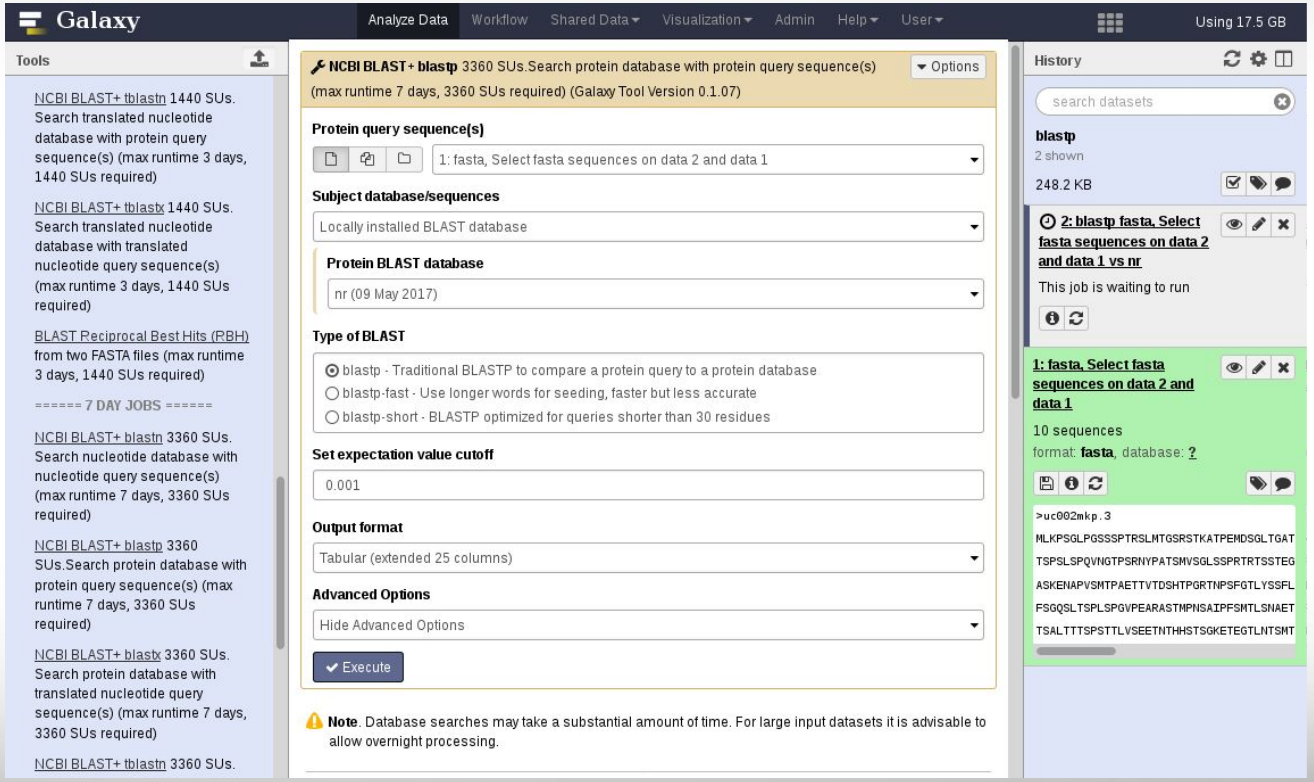


divided history items



Queued Jobs

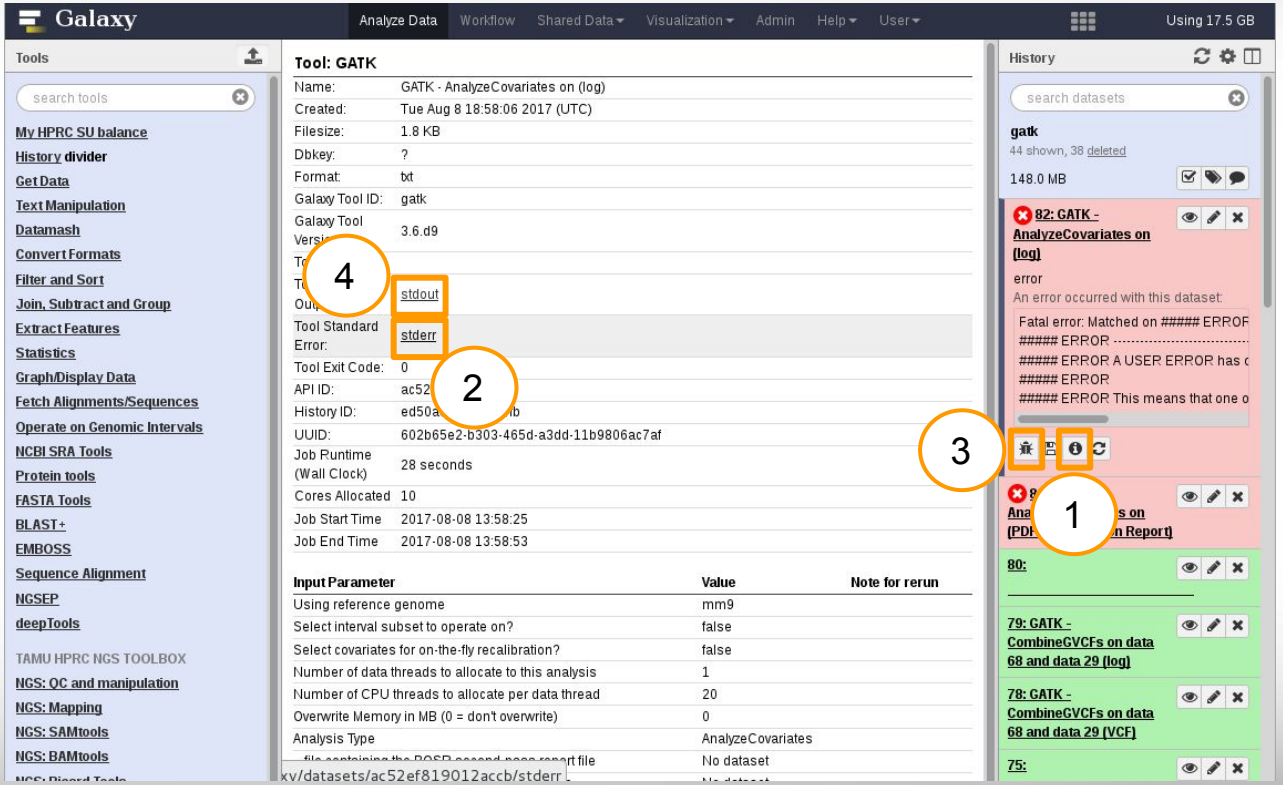
- Make sure you have enough SUs to run the job
  - My HPRC SU Balance
- Make sure your account is renewed for current fiscal year
  - My HPRC SU Balance
- Check to see if there is an Ada maintenance scheduled
- Check hprc.tamu.edu to see if Ada node usage is high >95%



Failed Jobs

- Check the stderr file link
- Check error log
- Read the email you received to see if job requires more memory or more time

  TERM_MEMLIMIT: job killed after reaching LSF memory usage limit.



Galaxy Notes on Ada

https://hprc.tamu.edu/wiki/index.php/Ada:Galaxy

Galaxy

Contents [hide]

Account Security

Do not share your Galaxy account with anyone. Galaxy uses the TAMU Central Authentication Service which is linked to your TAMU account.

Make sure you always logout of Galaxy by selecting User -> Logout and then click the Logout button on the next screen and then close your browser when you are finished using Galaxy.

FishCamp Galaxy Accounts

The FishCamp Galaxy instance is reserved for training purposes such as Galaxy workshops.

When requesting access to FishCamp for a training workshop, please include your ada NetID in your request.

- The FishCamp Galaxy is configured for training purposes.
  - Most jobs will run a maximum of 1 hour.
  - This is to enable jobs to be scheduled faster in the cluster queue.
  - Keep your input datasets small so that they will complete within one hour.

FishCamp Galaxy is not intended for research projects and data on FishCamp Galaxy should be considered to have short term accessibility.

Request a Maroon Galaxy account only if you have data to analyze.

The tools available on FishCamp and Maroon are the same.

usegalaxy.org Exercise

1. Add fasta file for S. cerevisiae chromosome I from genome.ucsc.edu to your current history and set the “Database/Build” attribute to the proper genome assembly
2. Run the “Compute Sequence Length” tool to get length of chromosome I
3. Add fasta file for chromosome M from genome.ucsc.edu to your history
4. Concatenate the two files into one file
5. Run the “Compute Sequence Length” tool on the concatenated file
6. Permanently delete the file you created in step 2.

Thank you.

Any question?