Getting data into Galaxy

Anne Pajon
Dave Clements
Maria Doyle
Saskia Hiltemann

Updated: PURL (Persistent URL): gxy.io/GTN:S00062

Plain-text slides

Tip: press P to view the presenter notes | Use arrow keys to move between slides

Questions

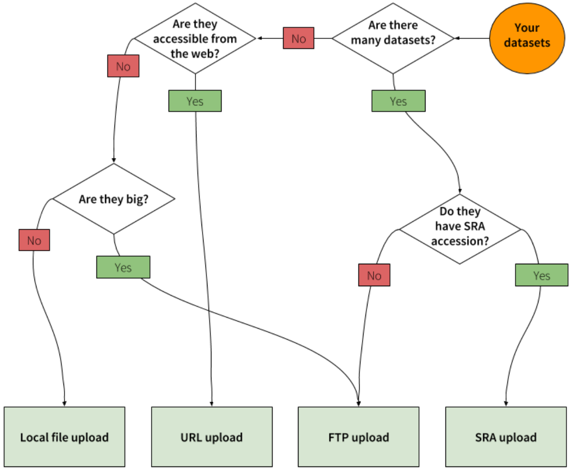
- How do I get my data into Galaxy?
- How do I get public data into Galaxy?

Getting data into Galaxy

Many ways to get data into your workspace

1. Import using **Get Data** sources e.g. UCSC, SRA
2. Import from a Galaxy **Data Library**
3. Import using **Upload File**
    - Import from your computer
    - Directly enter text
    - Import from a URL
    - Import using FTP
    - Import directly into Collection
    - Import using Rule Builder

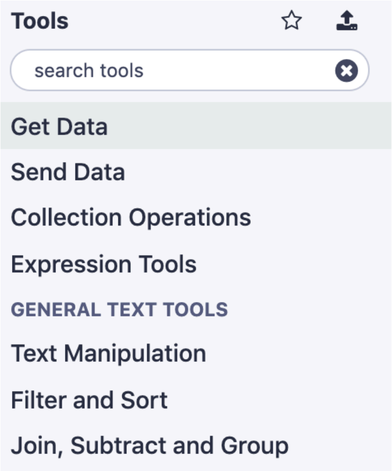
Best method depends on where the data is, and how big it is



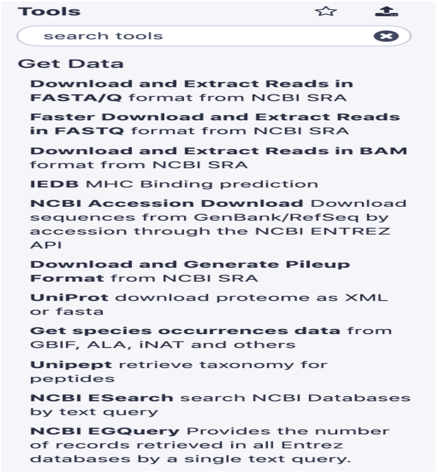
Source: Galaxy Community Hub

1. The *Get Data* toolbox section

- *Click* on the **Get Data** toolbox in the toolbox (the left panel)



- Expands to show data sources
  - E.g. UCSC, NCBI, Uniprot, ..
  - The specific data sources available on your Galaxy instance are determine by the server's administrator
- All of these data sources can bring datasets (files) into your Galaxy workspace (history)



Tools

search tools

Get Data

Download and Extract Reads in FASTA/Q format from NCBI SRA

Faster Download and Extract Reads in FASTQ format from NCBI SRA

Download and Extract Reads in BAM format from NCBI SRA

IEDB MHC Binding prediction

NCBI Accession Download Download sequences from GenBank/RefSeq by accession through the NCBI ENTREZ API

Download and Generate Pileup Format from NCBI SRA

UniProt download proteome as XML or fasta

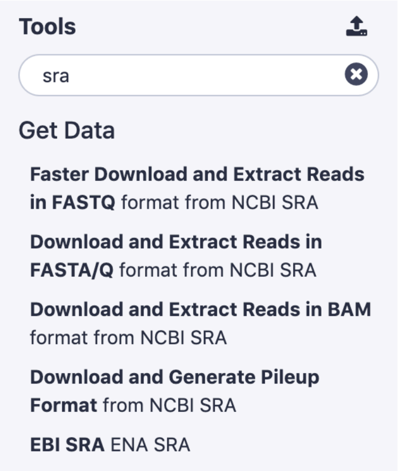
Get species occurrences data from GBIF, ALA, iNAT and others

Unipept retrieve taxonomy for peptides

NCBI ESearch search NCBI Databases by text query

NCBI EGQuery Provides the number of records retrieved in all Entrez databases by a single text query.

Two large data sources you can access through Galaxy are UCSC and SRA



Tools

sra

Get Data

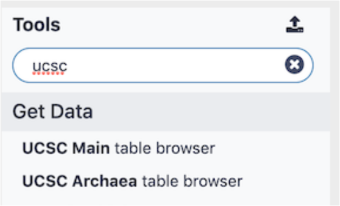
Faster Download and Extract Reads in FASTQ format from NCBI SRA

Download and Extract Reads in FASTA/Q format from NCBI SRA

Download and Extract Reads in BAM format from NCBI SRA

Download and Generate Pileup Format from NCBI SRA

EBI SRA ENA SRA



Tools

ucsc

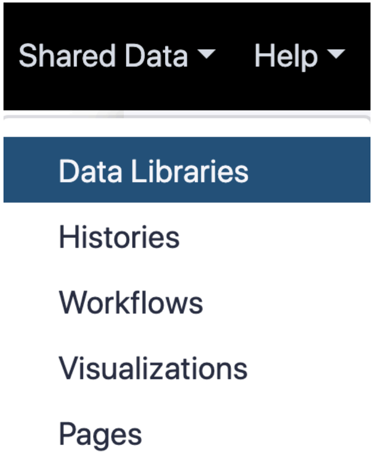
Get Data

UCSC Main table browser

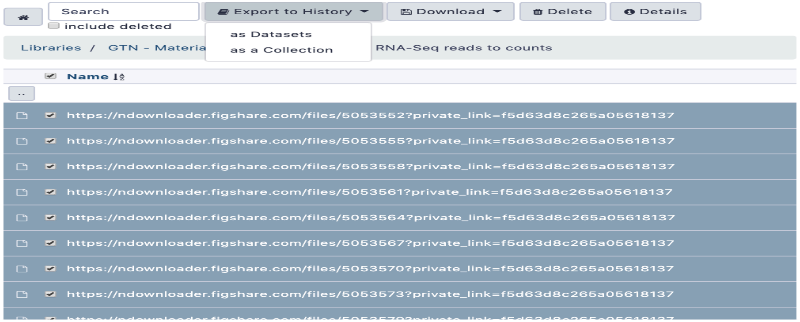
UCSC Archaea table browser

2. Import from Shared Data Library

- Top menu bar -> Shared Data -> Data Libraries
- Configured by a Galaxy Administator
- Can be imported directly into your history
- Example: all GTN tutorial data

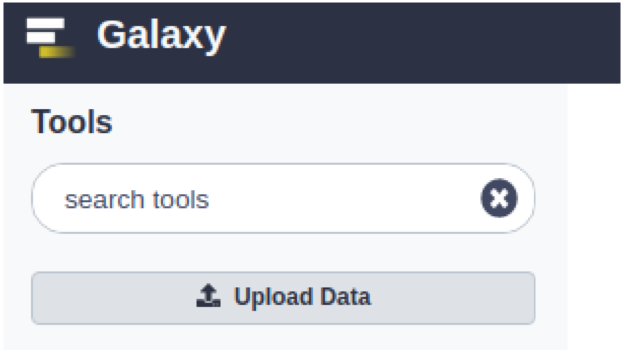


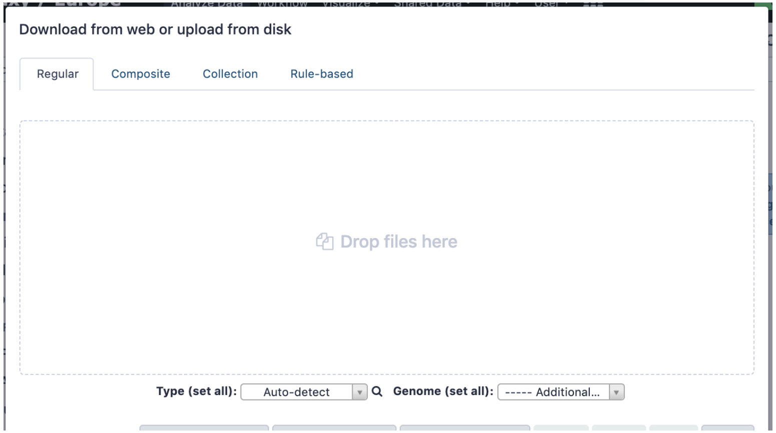
You can select the files you want and send to History as datasets or collection



☑ https://ndownloader.figshare.com/files/5053552?private_link=f5d63d8c265a05618137
☑ https://ndownloader.figshare.com/files/5053555?private_link=f5d63d8c265a05618137
☑ https://ndownloader.figshare.com/files/5053558?private_link=f5d63d8c265a05618137
☑ https://ndownloader.figshare.com/files/5053561?private_link=f5d63d8c265a05618137
☑ https://ndownloader.figshare.com/files/5053564?private_link=f5d63d8c265a05618137
☑ https://ndownloader.figshare.com/files/5053567?private_link=f5d63d8c265a05618137
☑ https://ndownloader.figshare.com/files/5053570?private_link=f5d63d8c265a05618137
☑ https://ndownloader.figshare.com/files/5053573?private_link=f5d63d8c265a05618137

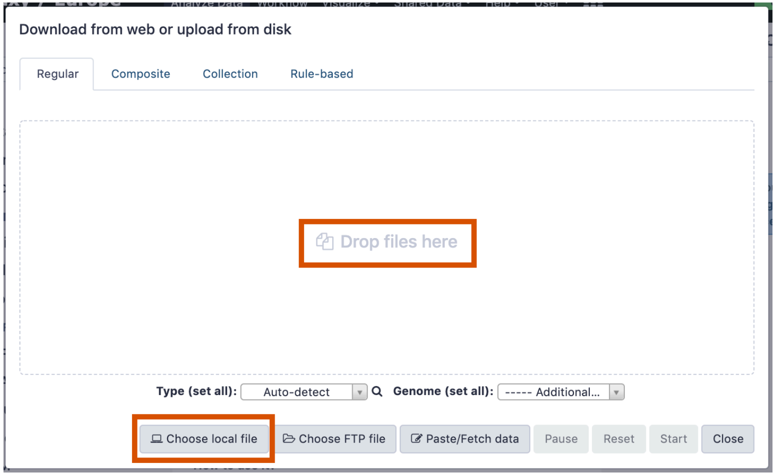
3. Upload from your computer





Type (set all): Auto-detect **Genome (set all):** ----- Additional...

Choose files



Download from web or upload from disk

Set Metadata

- **Datatype** (e.g. FastQ, VCF, BAM, tabular, ..)
    - Galaxy will autodetect by default (sometimes guesses wrong)
- **Genome Build** (e.g. hg19, mm9, ..)
    - must be set manually (can be done later as well)

- Can be set for all files at once:

- Or per file:

Start upload process:

- Once everything is ready, click the **Start** button

You can then close the form

All the items will appear in your history

and are ready to use when green.

Directly enter text

- Sometimes it's useful to file content directly.
  - only works if your dataset is tiny
  - choose **Paste/Fetch data**

Enter the data by typing (or pasting) it in the input box:

You can also set the datatype and build. *Click* **Start**, and then **Close**, and the new item shows up as **Pasted Entry** in your history.

Import using URL

The data might already be available on a web server somewhere. To avoid downloading data to your computer and uploading to Galaxy in two steps, you can instruct Galaxy to directly fetch the data from a given URL.

Select **Paste/Fetch data**

Enter the URLs (one per line) into the input box:

Click **Start**, and then **Close**, and the new items show up in your history with the URL as their name.

Import using FTP

- Why use FTP?
  - Older Galaxies did not support uploading files larger than 2GB in size
  - Many people are very comfortable using FTP to upload large datasets and you can sometimes resume interrupted uploads.
- How to use FTP
  - The Galaxy server's administrator must have enabled FTP on the server
  - You will need to create an account on that Galaxy Server
  - You will need to install FTP software, or to run FTP from the shell
  - See https://galaxyproject.org/ftp-upload/

Make sure you have an FTP client installed

- FileZilla is a free FTP client that is available on Windows, MacOS, and Linux
- There are many other options
- If you don't already have an FTP client, download and install FileZilla.

Establish FTP connection to your Galaxy server

- Provide
  - the instance's FTP server name (e.g. usegalaxy.org, ftp.usegalaxy.eu)
  - your full **username** (usually an email address) and **password**

Successfully connected

Right click on the files and upload them.

FTP Transfer in progress...

… and transfer complete.

Where did my files go?

- File Upload menu -> **Choose FTP files**

- Select files to import into your history
- Click **Start**

Import directly into Collection

- Select **Collection** tab at top of upload menu
- Add files as before (upload from computer, paste/fetch, FTP)

- Choose collection type (at bottom)
- Set metadata (file type, genome build)
- Click "Build"

- Name your collection
- Click **Create** button

- Collection is now imported in your history
- Click on it to expand it and view all files in collection

Import using Rule Based uploader

- When you want to import many files from URLs or Accession IDs directly into collection(s)
- Supports advanced "rules" for creating collections from sample sheets
- Click **Rule-based** tab at top of file upload window

Import using Rule Based uploader

Learn how to use it in the dedicated Rule Based Uploader tutorial

Thank You!

This material is the result of a collaborative work. Thanks to the Galaxy Training Network and all the contributors!

Authors: Anne Pajon Dave Clements Maria Doyle Saskia Hiltemann