Galaxy Project

Galaxy is an open source, web-based platform for data intensive biomedical research.

Core Values

Accessibility: Users without programming experience can easily upload/retrieve data, run complex tools and workflows, and visualize data

Reproducibility: Galaxy captures information so that any user can understand and repeat a complete computational analysis

Transparency: Users can share or publish their analysis

Loading Data

Importing Data:

- Copy/paste some text
- Upload files from your local computer
- Upload data from an internet URL
- Upload data from online databases: UCSC, BioMart, ENCODE, modENCODE, Flymine etc.
- Import from shared data (libraries, histories, pages)

Datatypes:

- Tools only accept input datasets with the appropriate datatypes
- When uploading a dataset, its datatype can be either automatically detected or assigned by the user
- Datasets produced by a tool have their datatype assigned by the tool

Sharing data

Share everything you do in Galaxy - histories, workflows, and visualizations

- Directly using a Galaxy account's email addresses on the same instance
- Using a web link, with anyone who knows the link
- Using a web link and publishing it to make it accessible to everyone from the *Shared Data* menu

Genome Assembly

Definitions:

Contig: a contiguous sequence in an assembly. A contig does not contain long stretches of unknown sequences (aka assembly **gaps**).

Scaffold: a sequence consists of one or multiple contigs connected by assembly gaps of typically inexact sizes. A scaffold is also called a **supercontig**, though this terminology is rarely used nowadays.

Assembly: a set of contigs or scaffolds.

Bioinformatics data format

FASTA: a text-based format for representing either nucleotide sequences or amino acid (protein) sequences, in which nucleotides or amino acids are represented using single-letter codes.

FASTQ: a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores (Phred). Both the sequence letter and quality score are each encoded with a single ASCII character for brevity. It's the standard sequencing output for Illumina and MGI sequencers.

SAM (Sequence Alignment Map): a text-based format originally for storing biological sequences aligned to a reference sequence developed by Heng Li and Bob Handsaker et al.

BAM (Binary Alignment Map): the comprehensive raw data of genome sequencing; it consists of the lossless, compressed binary representation of the SAM format. It's the standard sequencing output for PacBio sequencers.

CRAM (Compressed Reference-oriented Alignment Map): a compressed columnar file format for storing biological sequences aligned to a reference sequence.

Different types of input data

- Short reads (Illumina): numerous 👍, high quality 👍, cheap 👍, short 👎
- Long reads (PacBio, Nanopore): longer 👍, fewer 👎, (many) more errors 👎

Genome Assembly can be done with:

- only short reads
- only long reads
- both (hybrid assembly)

Specific algorithms for each

Genome Assembly algorithms

Detect overlaps between reads to build the longest possible sequences

Algorithms use graphs to represent overlapping reads/words

Two steps:

- Build a (huge) graph while reading the input data
- Try to find the longest paths traversing the graph

Two main types of algorithms:

- Short reads: de Bruijn Graphs
- Long reads: OLC (Overlap Layout Consensus)

What to do with my reads?

- Short reads => DBG assemblers (e.g. Spades, ABySS, DISCOVAR, Velvet, ...)
- Long reads => OLC assemblers (e.g Canu, Falcon, Hgap4, ...)
- Short + Long reads => hybrid assemblers (e.g. Unicycler, ...)

Hybrid assembly: long reads to resolve repeats, short reads to correct errors

Other data and tools for polishing (scaffolding, gap filling, ...)

Steps using Galaxy for project:

Step 1:

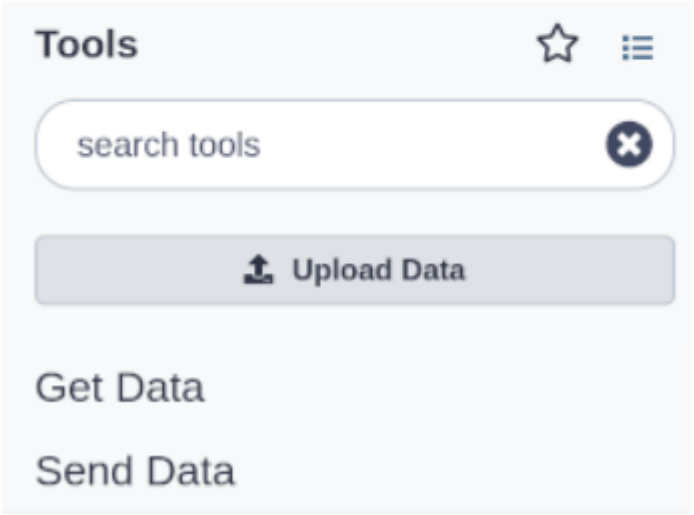
Upload the short_reads fasta file:

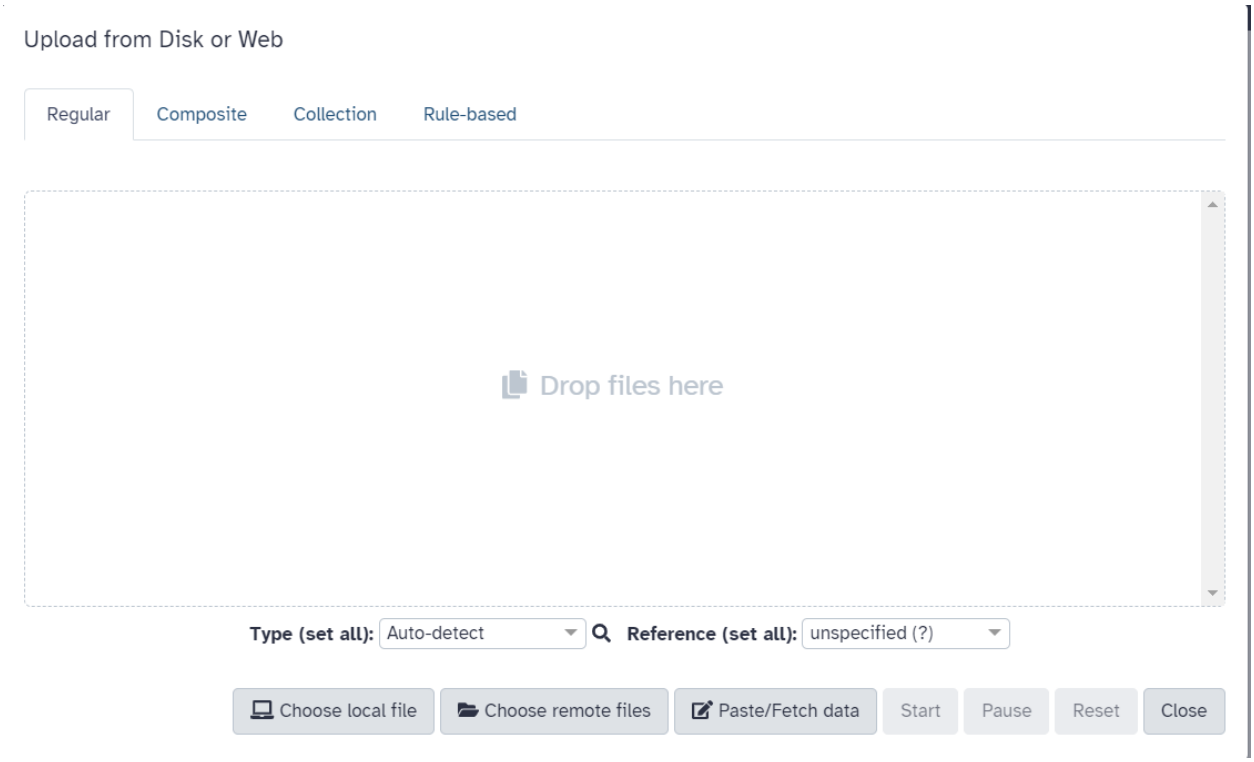
Upload a file

Your “Tools” are in the panel at the left.

Hands-on: Upload a file from URL

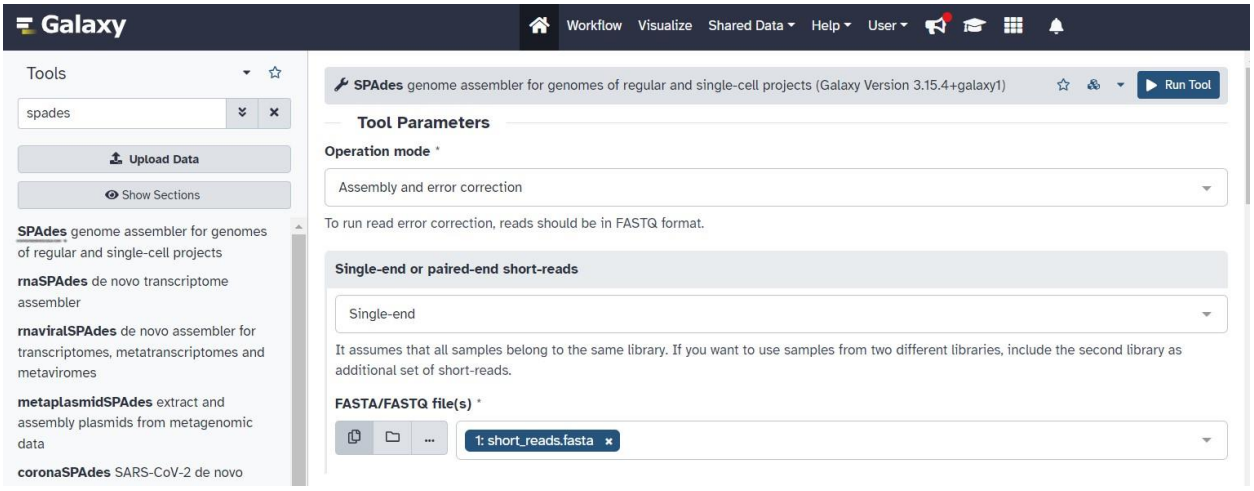
1. At the top of the **Tools** panel (on the left), click **Upload**





Step 2:

Using Galaxy to assemble genome
SPAdes genome assembler



At first SPAdes did not work when using blast
I tried MEGAHIT too .
but both works with a little difference in assembly result.