FungiDB: Data analysis via EuPathDB Galaxy

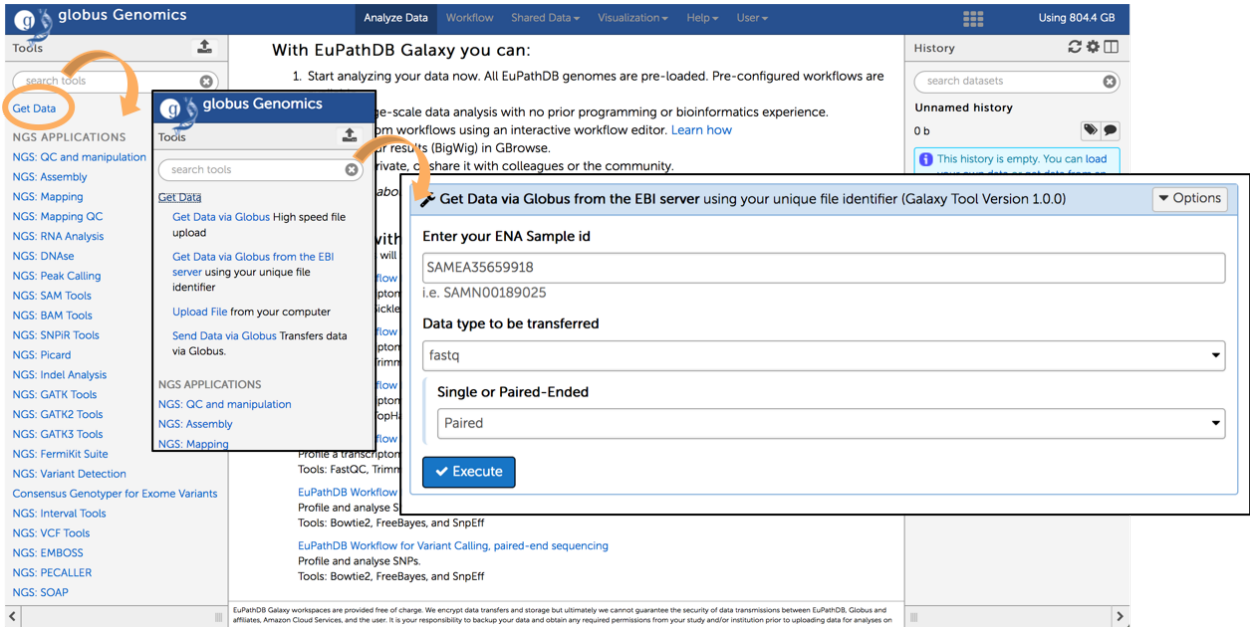
Variant Calling, Part I: Uploading data and starting the workflow (Group Exercise)

In this exercise we will work in groups to retrieve DNA sequence data from the sequence repository and analyze it for variants using a workflow in EuPathDB Galaxy.

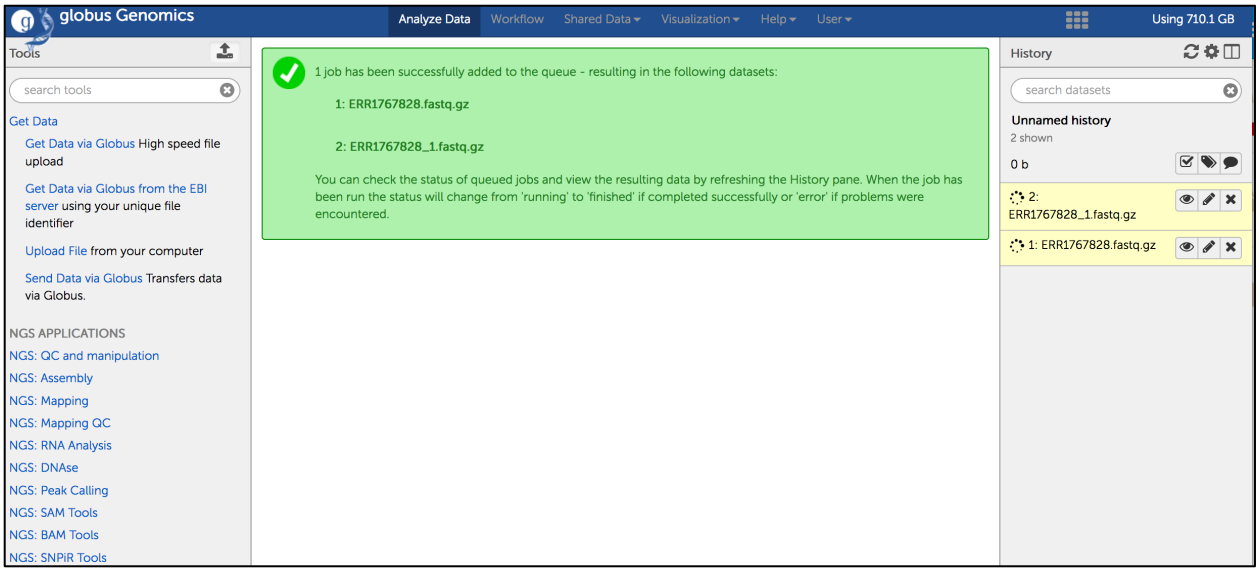
There are different ways to get data into Galaxy. Here we will use the sample ID and get the data using the "Get Data via Globus from the EBI server using your unique file identifier" link as well direct urls, as we have done this in the Galaxy RNA-Seq section.
Follow these steps to "Get Data via Globus from the EBI server":

1. Click on the *Get Data* link.
2. Click on the *Get Data via Globus from the EBI server* link.
3. The next window allows you to enter the sample ID. This ID starts with the letters 'SAM'. Choose the sample ID for your group from the list below and use it in this form.

   Note: it is very important that you select whether the data is single or paired-ended.
4. Once the form is properly filled, click on the *Execute* button to start the data transfer process.



If you click on the "Upload File from your computer" you will be able to use url links to initiate file download.

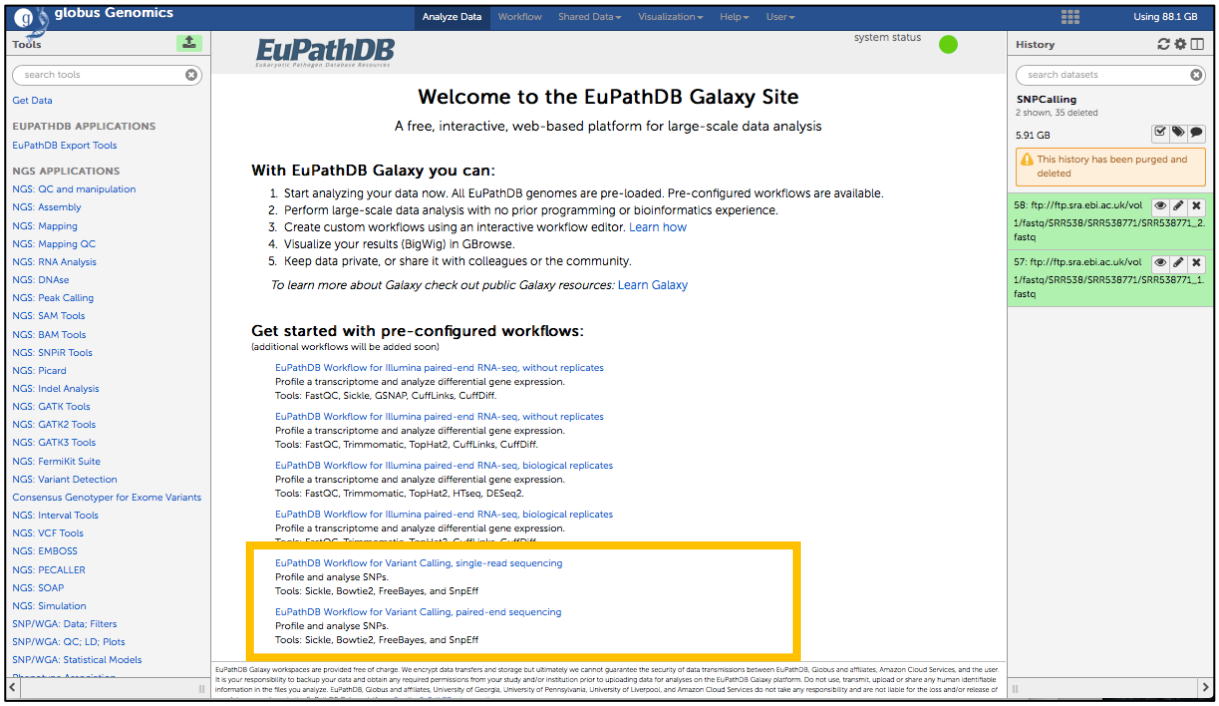


Group assignments:

Group assignments will be given in class and will be also available after the course online.

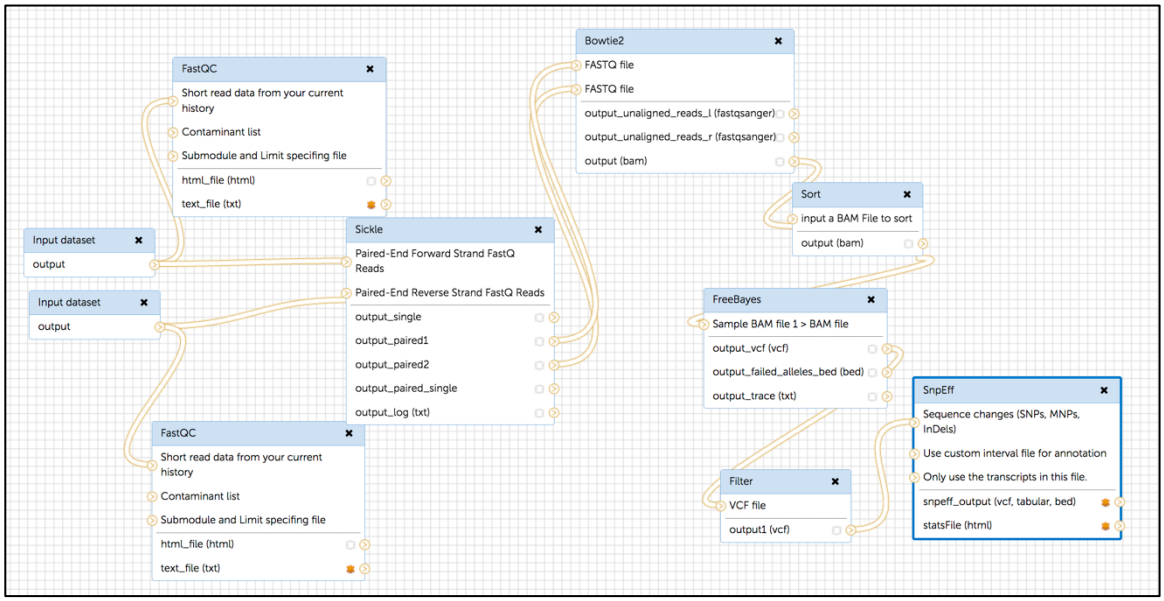
Running a variant calling workflow:

- Once the data files have been transferred into your galaxy history you need to choose an appropriate workflow. EuPathDB provides some preconfigured workflows on the EuPathDB Galaxy instance home page.
- Remember to choose the appropriate workflow – Single-read or paired ended.

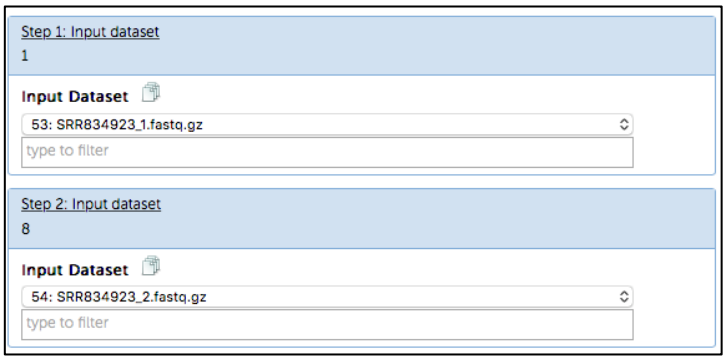


The pre-configured workflows follow these steps:

- Determine quality of the reads in your files and generates FASTQC reports
- Trim reads based on their quality scores
- Align reads to a reference genome using Bowtie2 and generating coverage plots
- Sort alignments with respect to their chromosomal positions
- Detect variants using FreeBayes
- Filter SNP candidates
- Analyze and annotate of variants, and calculation of the effects via SnpEff



- Next, set workflow parameters.
  - Make sure that the input steps for paired-end are set to the *xxxx_1.fastq.gz* and *xxxx_2.fastq.gz* as by default both have the same one selected.



- Select the correct reference genome (for steps: Bowtie2, FreeBayes, SnpEff)
- Click on the *Run Workflow* button.