Variant calling in VEuPathDB galaxy (Part 1)

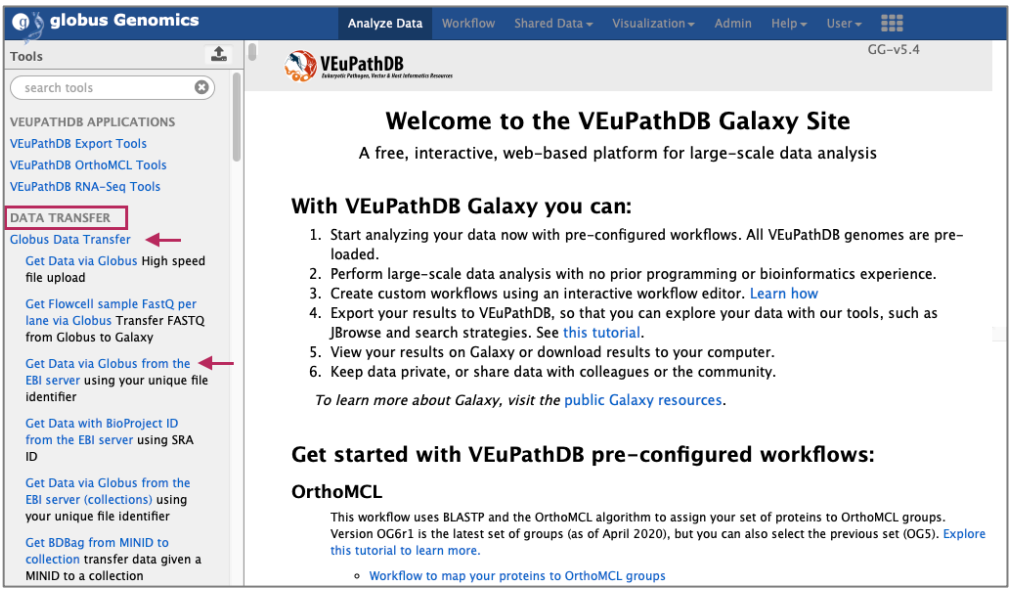
Learning objectives:

1. Retrieve DNA sequence data from the sequence repository EBI and upload data to VEuPathDB Galaxy using Globus Data Transfer;
2. Name a new project/history;
3. Deploy a Variant calling workflow in the VEuPathDB Galaxy.

There are different ways to get data into Galaxy. In this exercise we will use Globus Data Transfer to get data from the EBI server using a unique project ID.

1. Retrieve DNA sequence data from the sequence repository and upload data to VEuPathDB Galaxy using Globus Data Transfer option.

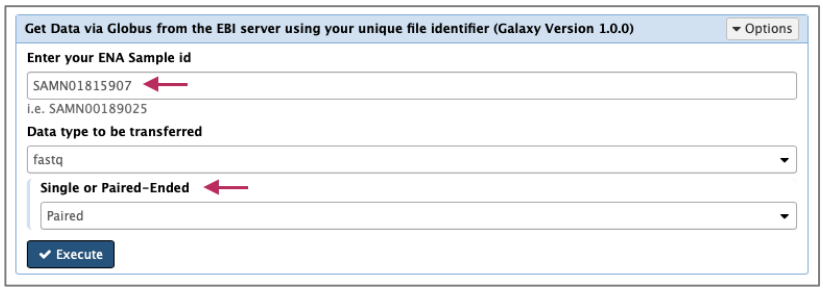
a. Click on the “Globus Data Transfer” menu on the left to expand the Data Transfer section.
b. Click on the “Get Data via Globus from the EBI server” link.



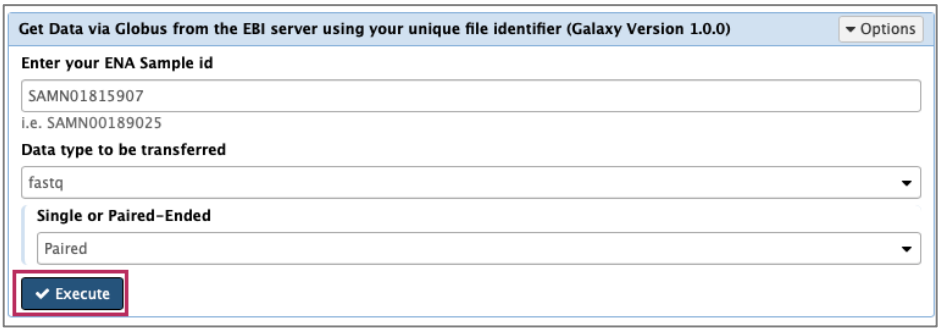
c. Enter ENA sample ID and define the dataset type to be transferred into the VEuPathDB Galaxy workspace.

The ENA ID should start with the letters ‘SAM’. For this exercise, we will use SAMN01815907, which is a paired-ended dataset. Take care to specify

whether a dataset is a single or paired-ended as incorrect selection will cause the upload to fail.

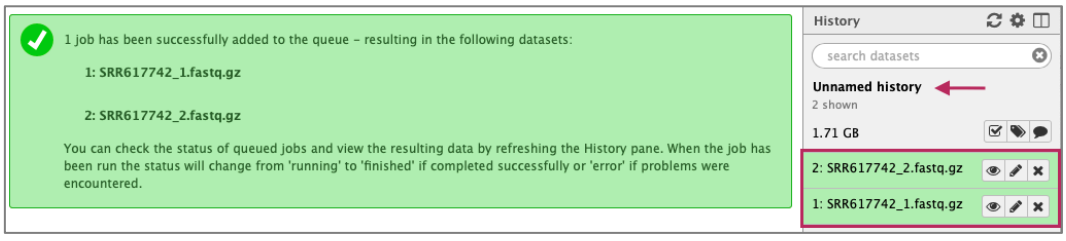


d. Once the form is properly filled, click on the “Execute” button to start the data transfer process.



e. When the job has been successfully deployed and added to the queue, the screen will refresh, and the added job will appear in the history on the right.

Note: new jobs are highlighted in grey, in progress – yellow, and those completed are in green.

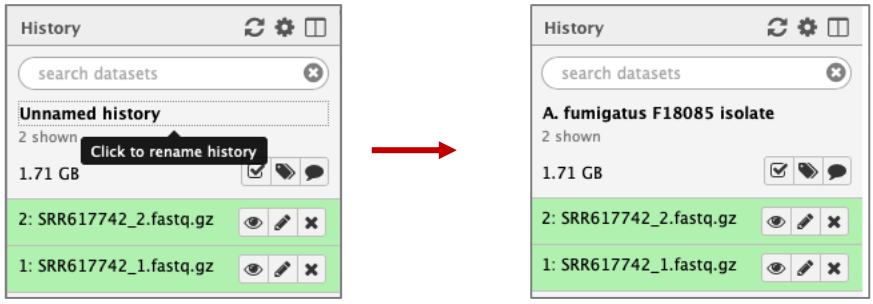


Notice that there are two files appearing in the history on the right. This is because the uploaded data is paired-ended.

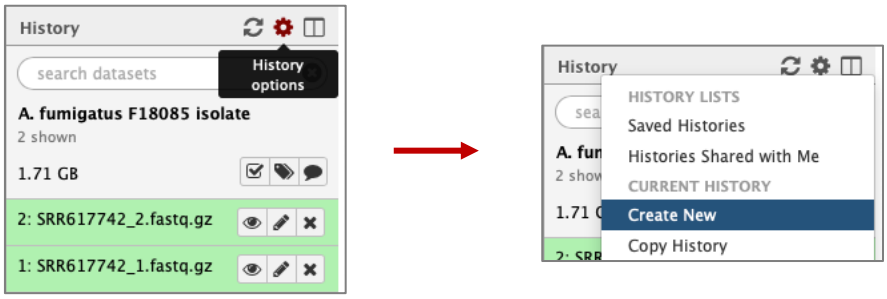
2. Rename your history.

By default, all new jobs will be added to the current history on the right. Unless renamed, the history will show up in your history as “Unnamed history”. Let’s rename the history to help us track this project in the future.

a. Click on the “Unnamed history” and type “A. fumigatus F18085 isolate”, and then press “enter” to rename this history.



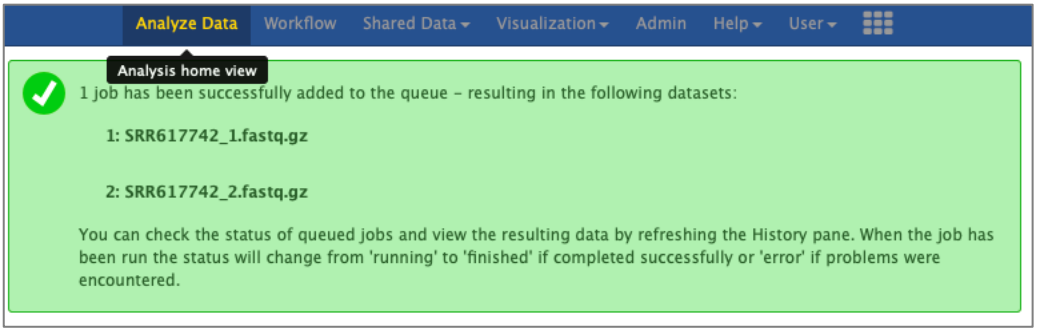
Note: if you would like to start a new project/history, click on the wheel button at the top of the history section and select “Create a new history”.



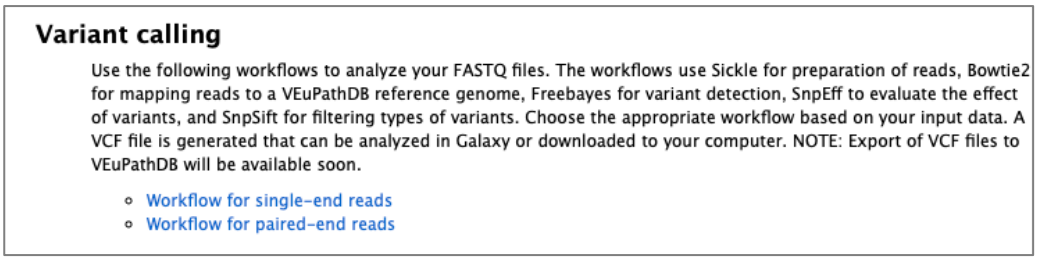
3. Deploy a Variant calling workflow.

VEuPathDB Galaxy main landing page has several workflows for variant calling.

a. To navigate to the main page, click on the “Analyze Data”, which is located in the main menu at the top.

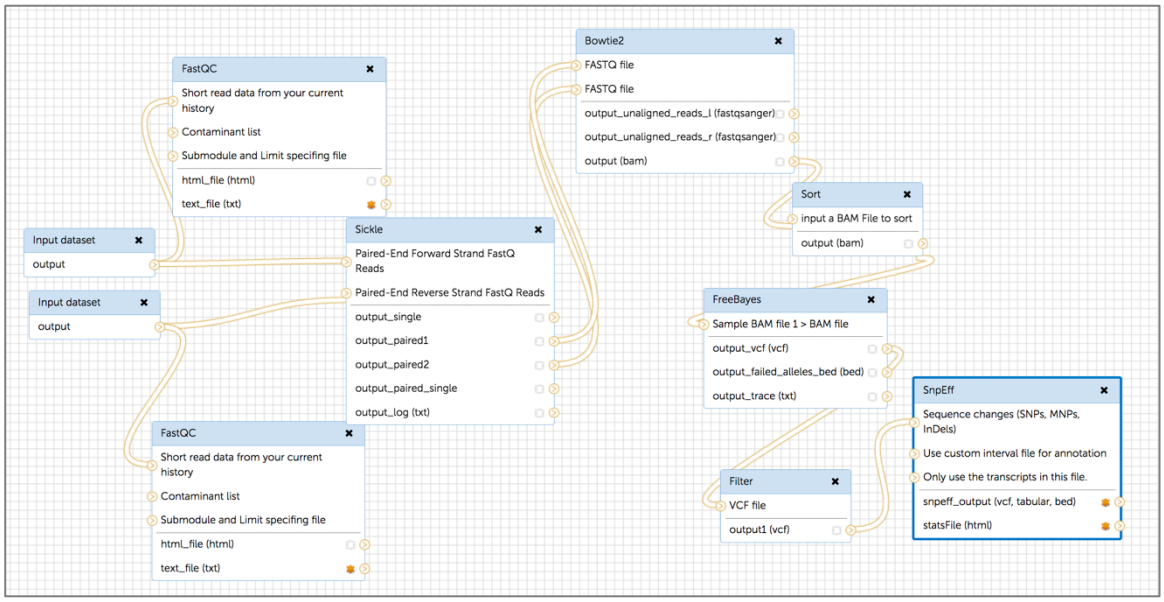


b. Scroll down to the Variant calling section and choose the workflow for paired-end reads.



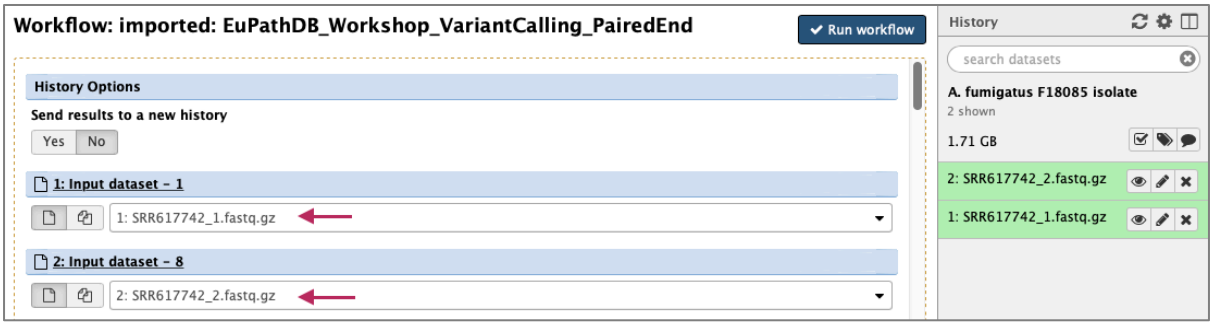
The pre-configured variant calling workflows include the following steps:

- Determine quality of the reads and generate reports (FastQC);
- Trim reads based on their quality scores (Sickle);
- Align reads to a reference genome using Bowtie2 and generate coverage plots ;
- Sort alignments with respect to their chromosomal positions (Sort);
- Detect variants (FreeBayes);
- Filter SNP candidates (Filter);
- Analyze and annotate variants, and calculate the effects of SNPs via SnpEff.

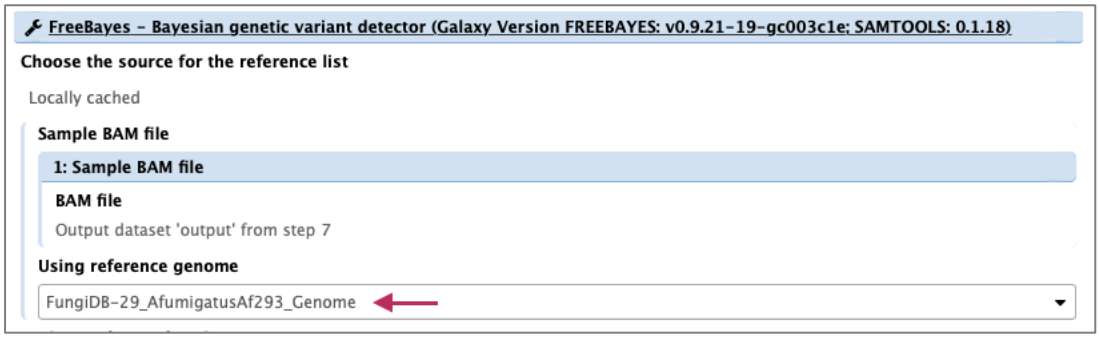


c. Click on the workflow for paired-end reads and set workflow parameters.

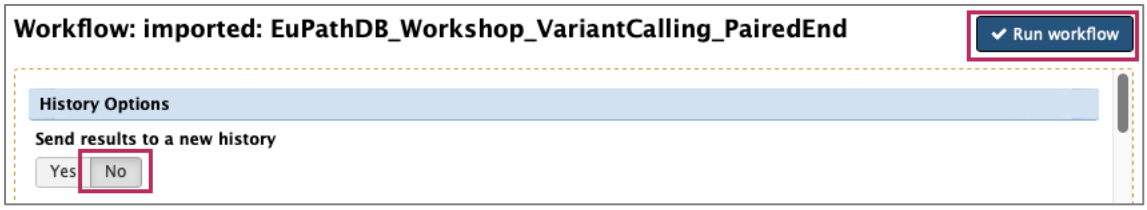
- Make sure that the input steps for paired-end data are set to the xxxx_1.fastq.gz and xxxx_2.fastq.gz file (by default the same file will be selected in both fileds).



- Select the correct reference genome.
    - Select *Aspergillus fumigatus* Af293 as a reference genome (steps: Bowtie2, FreeBayes, SnpEff).



- Choose to deploy the analysis within the same history and click on the Run workflow button.



Note: You can use the same workflow to analyze multiple samples in batches. The. Upload steps remain the same, however, when setting up the workflow, click on multiple dataset button within the input dataset section.