

Variant Calling using EuPathDB Galaxy

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. For this workshop we will use the workshop specific galaxy site:

<https://eupathdbworkshop.globusgenomics.org/>

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. Follow these steps:

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-end.
4. Once the form is properly filled, click on the ‘Execute’ button to start the data transfer process.

The screenshot displays the EuPathDB Galaxy interface. On the left, a sidebar lists various tools under 'NGS APPLICATIONS', with 'Get Data' circled in red. A red arrow points from this link to a secondary window titled 'globus Genomics' which lists data acquisition options. Another red arrow points from the option 'Get Data via Globus from the EBI server using your unique file identifier' to a configuration form. The form, titled 'Get Data via Globus from the EBI server using your unique file identifier (Galaxy Tool Version 1.0.0)', contains the following fields: 'Enter your ENA Sample id' with the value 'SAMEA35659918', 'Data type to be transferred' set to 'fastq', and 'Single or Paired-Ended' set to 'Paired'. An 'Execute' button is located at the bottom of the form. The background shows the main Galaxy workspace with a 'Tools' panel and a 'History' panel.

Groups:

Group 1: *Plasmodium berghei* wild type

Sample ID: SAMN04386828

<https://www.ebi.ac.uk/ena/data/view/SAMN04386828>

Group 2: *Plasmodium berghei* drug resistant mutant

Sample ID: SAMN04386825

<https://www.ebi.ac.uk/ena/data/view/SAMN04386825>

Group 3: *Cryptosporidium* field isolate (clinic visit sample)

Sample ID: SAMEA104459068

<https://www.ebi.ac.uk/ena/data/view/SAMEA104459068>

Group 4: *Cryptosporidium* field isolate (Diarrheal sample)

Sample ID: SAMEA104459070

<https://www.ebi.ac.uk/ena/data/view/SAMEA104459070>

Group 5: *Toxoplasma gondii* RH parental strain (type I strain)

Sample ID: SAMN06112744

<http://www.ebi.ac.uk/ena/data/view/SAMN06112744>

Group 6: *Toxoplasma gondii* RH IBET-151 resistant mutant (type I strain)

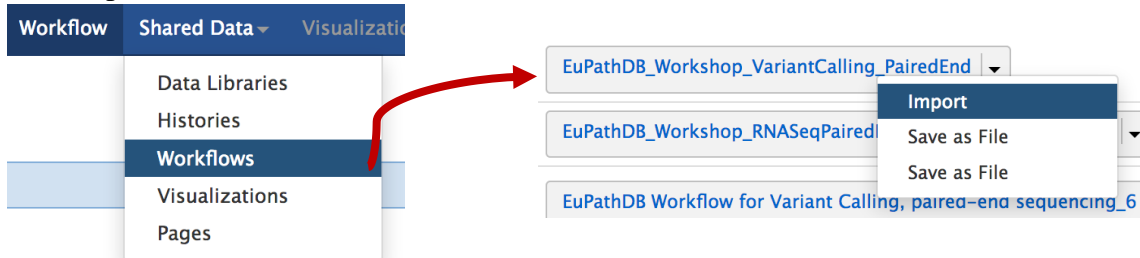
Sample ID: SAMN06112745

<http://www.ebi.ac.uk/ena/data/view/SAMN06112745>

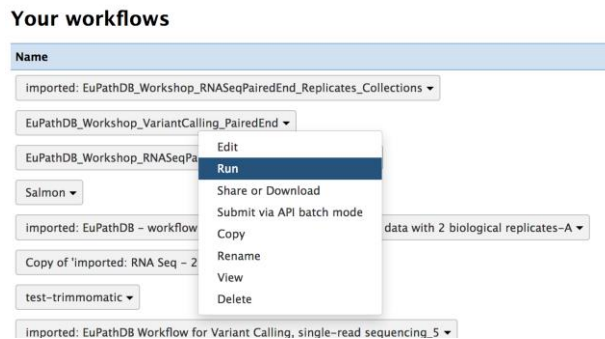
The screenshot displays the Globus Genomics web interface. At the top, the navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main content area features a green notification box with a checkmark icon, stating: '1 job has been successfully added to the queue - resulting in the following datasets: 1: ERR1767828.fastq.gz 2: ERR1767828_1.fastq.gz'. Below this, a message reads: 'You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.' On the left sidebar, there are sections for 'Tools' (with a search bar), 'Get Data' (with options like 'Get Data via Globus High speed file upload', 'Get Data via Globus from the EBI server using your unique file identifier', 'Upload File from your computer', and 'Send Data via Globus Transfers data via Globus'), and 'NGS APPLICATIONS' (listing various tools like QC and manipulation, Assembly, Mapping, etc.). On the right, the 'History' panel shows 'Unnamed history' with '2 shown' items: 'ERR1767828_1.fastq.gz' and 'ERR1767828.fastq.gz', each with view, edit, and delete icons. The top right corner indicates 'Using 710.1 GB'.

Running a variant calling workflow:

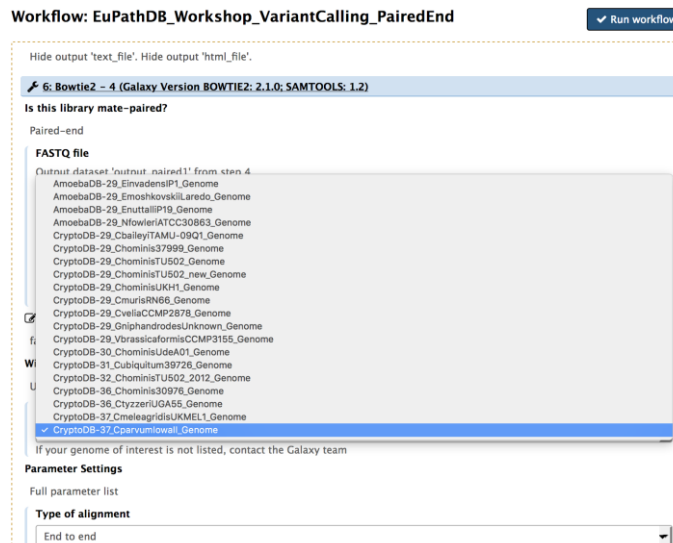
- a. Use the shared workflow called: `EuPathDB_Workshop_VariantCalling_PairedEnd`. To use a shared workflow, first you must import it by going to 'workflows' under the shared data tab, clicking on the desired workflow and selecting 'import' from the dropdown menu.



- b. To run this workflow, go to the workflow section, click on the workflow with the correct name and select "Run"



- c. Remember to select the correct reference genome. Check with the other group using a sample from the same experiment and make sure you both agree on which reference genome to use (Bowtie2, FreeBayes, SnpEff)



- d. Click on the 'Run Workflow' button.