

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'. A red circle highlights the 'Get Data' link in this sidebar. A red arrow points from this link to a central panel titled 'With EuPathDB Galaxy you can:' which lists several options. Another red arrow points from the 'Get Data via Globus from the EBI server using your unique file identifier' option to a detailed configuration window. This window is titled 'Get Data via Globus from the EBI server using your unique file identifier (Galaxy Tool Version 1.0.0)'. It contains the following fields and options:

- Enter your ENA Sample id:** A text input field containing 'SAMEA35659918'.
- Data type to be transferred:** A dropdown menu set to 'fastq'.
- Single or Paired-Ended:** A dropdown menu set to 'Paired'.
- Execute:** A blue button with a checkmark icon.

At the bottom of the page, a small disclaimer states: 'EuPathDB Galaxy workshops are provided free of charge. We encrypt data transfers and storage but ultimately we cannot guarantee the security of data transmissions between EuPathDB, Globus and affiliates, Amazon Cloud Services, and the user. It is your responsibility to backup your data and obtain any required permissions from your study and/or institution prior to uploading data for analyses on'.

Groups:

Group 1: *Plasmodium falciparum* drug resistant field isolate

Sample ID: SAMN01087919

<http://www.ebi.ac.uk/ena/data/view/SAMN01087919>

Group 2: *Babesia microti* field isolate (Rhode Island)

Sample ID: SAMEA3918179

<http://www.ebi.ac.uk/ena/data/view/SAMEA3918179>

Group 3: *Babesia microti* field isolate (Wisconsin)

Sample ID: SAMEA3918185

<http://www.ebi.ac.uk/ena/data/view/SAMEA3918185>

Group 4: *Candida albicans* CHN1

Sample ID: SAMN00974105

<http://www.ebi.ac.uk/ena/data/view/SAMN00974105>

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:'. Below this, two dataset identifiers are listed: '1: ERR1767828.fastq.gz' and '2: ERR1767828_1.fastq.gz'. A message below the list 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.' To the right, the 'History' panel shows a search bar and a list of datasets under the heading 'Unnamed history'. Two datasets are visible: '2: ERR1767828_1.fastq.gz' and '1: ERR1767828.fastq.gz', each with icons for viewing, editing, and deleting. The left sidebar contains 'Tools' and 'Get Data' options, along with a list of 'NGS APPLICATIONS' such as 'QC and manipulation', 'Assembly', 'Mapping', and 'Peak Calling'.

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 ended or paired ended.

The screenshot shows the EuPathDB Galaxy interface. The main content area is titled "With EuPathDB Galaxy you can:" and lists five points: 1. Start analyzing your data now. All EuPathDB genomes are pre-loaded. Pre-configured workflows are available. 2. Perform large-scale data analysis with no prior programming or bioinformatics experience. 3. Create custom workflows using an interactive workflow editor. 4. Visualize your results (BigWig) in GBrowse. 5. Keep data private, or share it with colleagues or the community. Below this is a link to "Learn more about Galaxy check out public Galaxy resources: Learn Galaxy".

Under the heading "Get started with pre-configured workflows:", there are four workflow options. The first two are for RNA-seq analysis. The last two are for variant calling and are highlighted with a red box:

- EuPathDB Workflow for Variant Calling, single-read sequencing**
Profile and analyse SNPs.
Tools: Bowtie2, FreeBayes, and SnpEff
- EuPathDB Workflow for Variant Calling, paired-end sequencing**
Profile and analyse SNPs.
Tools: Bowtie2, FreeBayes, and SnpEff

The left sidebar shows a "Tools" section with a search bar and a list of NGS applications including QC and manipulation, Assembly, Mapping, Mapping QC, RNA Analysis, DNase, Peak Calling, SAM Tools, BAM Tools, SNP Tools, Picard, Indel Analysis, GATK Tools, GATK2 Tools, GATK3 Tools, FermiKit Suite, Variant Detection, Consensus Genotyper for Exome Variants, Interval Tools, VCF Tools, EMBOSS, PICALER, and SOAP.

The right sidebar shows a "History" section with a search bar and a message: "This history is empty. You can load your own data or get data from an external source".

- Set workflow parameters. **Note that the trimming step “Sickle” has a parameter to select the “quality type”. The default is often “Illumina”. This will not work and has to be changed to Sanger.**
- Select the correct reference genome (Bowtie2, FreeBayes, SnpEff)
- Click on the ‘Run Workflow’ button.

The screenshot shows the "Step 3: Sickle (version SICKLE: 070113)" configuration screen. The workflow ID is 3. The "Single-End or Paired-End reads?" section is set to "Single-End". The "Single-End FastQ Reads" section shows "Output dataset 'output' from step 1". The "Quality type" dropdown menu is open, showing "Illumina" selected, with "Solexa" and "Sanger" as options. The "Length Threshold" is set to 20. The "Don't do 5' trimming" checkbox is checked, and the "Discard sequences with Ns" checkbox is also checked.