

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.

The screenshot displays the EuPathDB Galaxy web interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main content area is titled 'With EuPathDB Galaxy you can:' and lists various analysis options. A 'Tools' sidebar on the left is open, showing a search bar and a list of tool categories under 'Get Data'. The 'Get Data via Globus from the EBI server using your unique file identifier' tool is selected and its configuration form is displayed. The form includes the following fields and options:

- Enter your ENA Sample id:** A text input field containing 'SAMEA35659918' and a sub-label 'i.e. SAMN00189025'.
- 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.

The background shows a 'History' panel on the right with a search bar and a message: 'This history is empty. You can load...'. The bottom of the page contains a small disclaimer about data security and backups.

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

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 sub-message explains: '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 left is a 'Tools' sidebar with a search bar and various data acquisition options. To the right is a 'History' panel with a search bar and a list of datasets, including 'ERR1767828_1.fastq.gz' and 'ERR1767828.fastq.gz'.

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.

Welcome to the EuPathDB Galaxy Site
A free, interactive, web-based platform for large-scale data analysis

With EuPathDB Galaxy you can:

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. [Learn how](#)
4. Visualize your results (BigWig) in GBrowse.
5. Keep data private, or share it with colleagues or the community.

To learn more about Galaxy check out public Galaxy resources: [Learn Galaxy](#)

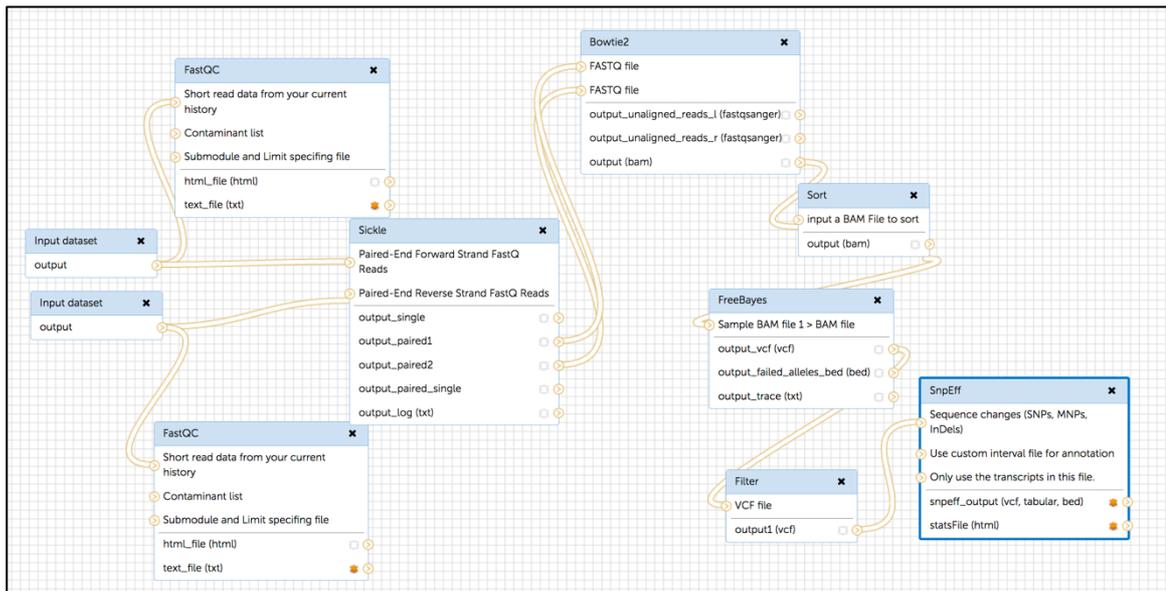
Get started with pre-configured workflows:
(additional workflows will be added soon)

- EuPathDB Workflow for Illumina paired-end RNA-seq, without replicates
Profile a transcriptome and analyze differential gene expression.
Tools: FastQC, Sickle, GSNAP, CuffLinks, CuffDiff.
- EuPathDB Workflow for Illumina paired-end RNA-seq, without replicates
Profile a transcriptome and analyze differential gene expression.
Tools: FastQC, Trimmomatic, TopHat2, CuffLinks, CuffDiff.
- EuPathDB Workflow for Illumina paired-end RNA-seq, biological replicates
Profile a transcriptome and analyze differential gene expression.
Tools: FastQC, Trimmomatic, TopHat2, HTSeq, DESeq2.
- EuPathDB Workflow for Illumina paired-end RNA-seq, biological replicates
Profile a transcriptome and analyze differential gene expression.
Tools: FastQC, Trimmomatic, TopHat2, HTSeq, DESeq2.
- EuPathDB Workflow for Variant Calling, single-read sequencing
Profile and analyse SNPs.
Tools: Sickle, Bowtie2, FreeBayes, and SnpEff
- EuPathDB Workflow for Variant Calling, paired-end sequencing
Profile and analyse SNPs.
Tools: Sickle, Bowtie2, FreeBayes, and SnpEff

EuPathDB Galaxy workflows 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 the EuPathDB Galaxy platform. Do not use, transmit, upload or share any human identifiable information in the files you analyze. EuPathDB, Globus and affiliates, University of Georgia, University of Pennsylvania, University of Liverpool, and Amazon Cloud Services do not take any responsibility and are not liable for the loss and/or release of

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.

Step 1: Input dataset
1

Input Dataset

53: SRR834923_1.fastq.gz

type to filter

Step 2: Input dataset
8

Input Dataset

54: SRR834923_2.fastq.gz

type to filter

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