

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 globus Genomics interface. On the left, a sidebar lists various NGS applications, with 'Get Data' circled in red. A red arrow points from this link to a secondary window titled 'globus Genomics' which lists several options for data acquisition. Another red arrow points from the option 'Get Data via Globus from the EBI server using your unique file identifier' to a larger configuration window. This configuration 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 interface, a small disclaimer states: 'EuPathDB Galaxy workspaces 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: 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, the 'Tools' sidebar lists various data acquisition methods and NGS applications. On the right, the 'History' panel shows a search bar and a list of datasets under 'Unnamed history', including 'ERR1767828_1.fastq.gz' and 'ERR1767828.fastq.gz'.

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.

globus Genomics

Analyze Data Workflow Shared Data Visualization Help User

Using 804.4 GB

Tools

search tools

Get Data

NGS APPLICATIONS

NGS: QC and manipulation

NGS: Assembly

NGS: Mapping

NGS: Mapping QC

NGS: RNA Analysis

NGS: DNase

NGS: Peak Calling

NGS: SAM Tools

NGS: BAM Tools

NGS: SNPir Tools

NGS: Picard

NGS: Indel Analysis

NGS: GATK Tools

NGS: GATK2 Tools

NGS: GATK3 Tools

NGS: FermiKit Suite

NGS: Variant Detection

Consensus Genotyper for Exome Variants

NGS: Interval Tools

NGS: VCF Tools

NGS: EMBOSS

NGS: PICALLER

NGS: SOAP

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, 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, CuffLinks, CuffDiff.

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

History

search datasets

Unnamed history

0 b

This history is empty. You can load your own data or get data from an external source

EuPathDB 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

- 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.

Step 3: Sickle (version SICKLE: 070113)
3

Single-End or Paired-End reads?

Single-End

Single-End FastQ Reads

Output dataset 'output' from step 1

Quality type

Illumina
Illumina
Solexa
Sanger

Length Threshold

20

Don't do 5' trimming

False

Discard sequences with Ns

False