What is Galaxy?

Galaxy is an open, web-based platform for data analysis under the FAIR principles of data sharing and re-use. Galaxy is an open-source platform that allows you to perform, reproduce, and share complete analyses without the use of command line scripting. The VEuPathDB project developed its own Galaxy instance in collaboration with Globus.

The VEuPathDB Galaxy offers pre-loaded genomes, pre-configured workflows and other tools for private data analysis and display. A custom-built set of tools also allows the ability to export Galaxy results into private workspaces within VEuPathDB sites (My Workspace > My data sets section). The datasets within the "My data sets" workspace can be explored using the FungiDB interface and tools and cross-referenced with the public data integrated in FungiDB.

VEuPathDB Galaxy access requires an account with FungiDB/VEuPathDB. The account is free and can be used to sign-in into any VEuPathDB genomics site.

The Galaxy instance is not meant for long term data storage. Datasets are automatically deleted after 60 days. To save your data, download your analysis results locally and then *delete and purge* files to free up space for your next analysis.

The Galaxy project offers extensive learning materials that can be accessed here: https://wiki.galaxyproject.org/Learn#Galaxy_101

Important: The Galaxy module consists of RNA-Seq and SNP analysis modules. These are concurrent sessions. This exercise will be carried out in groups of 4 people using the workshop Galaxy instance. Please do not use live FungiDB.org for this exercise. The detailed tutorials for both modules are available to all course participants.

Variant Calling analysis, Part I.

Learning objectives:

- Become familiar with the VEuPathDB Galaxy workspace.
- Upload raw data into Galaxy workspace and run a pre-configured SNP workflow

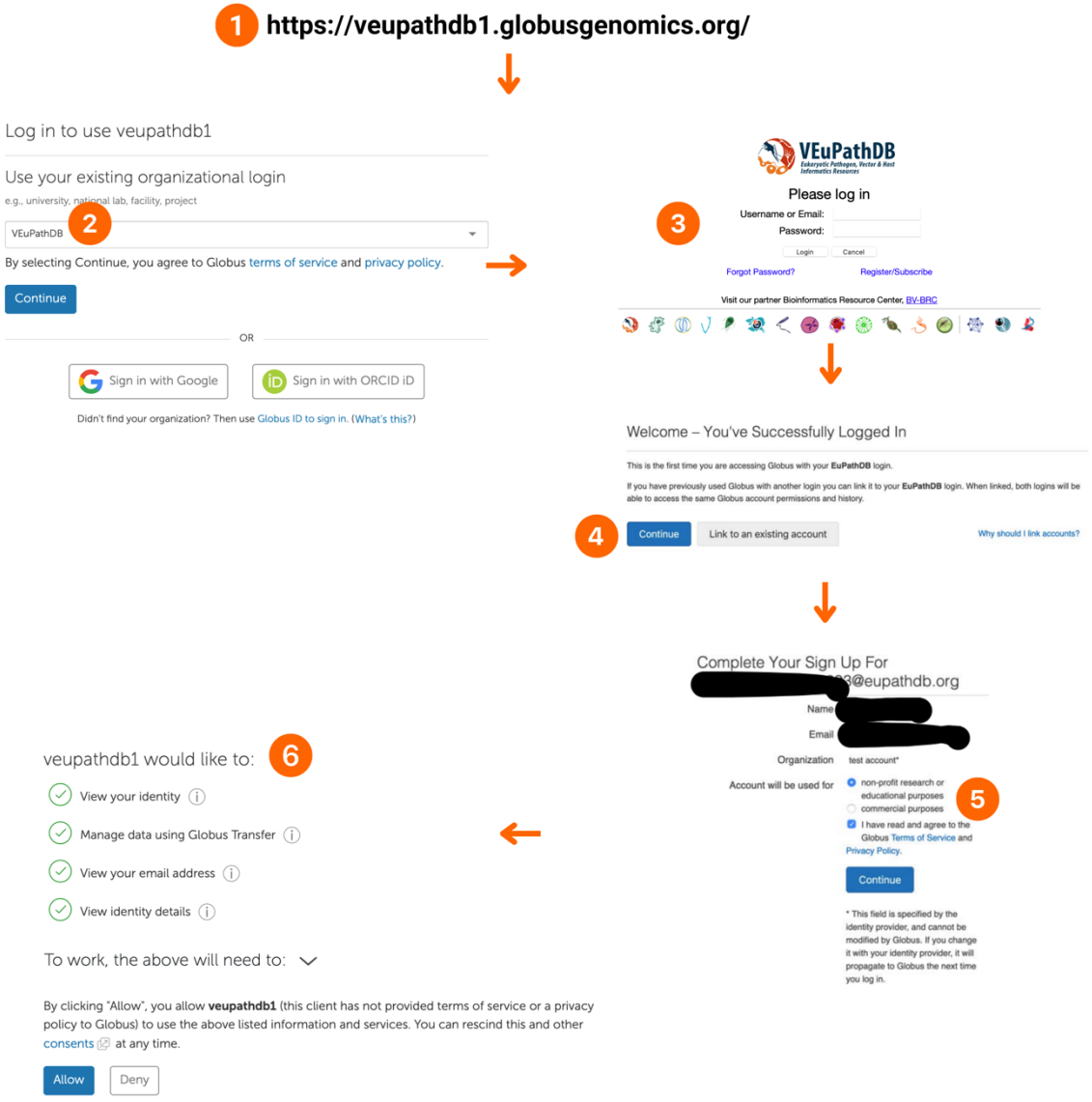
For this exercise, we will retrieve raw sequence files from the "shared history" section in VEuPathDB Galaxy and then run files through a pre-configured RNA-Seq workflow that will align the data to a reference genome, calculate expression values and determine differential expression.

Important: We will be working in groups of four people but only one person in each group should download data and deploy the pre-configured workflow. The other members' roles are to ensure that the correct datasets are used and that the correct workflow parameters are selected. In the Part 2 of this exercise, everyone will get a copy of the workflow output and practice how to perform data analysis.

- **Access the VEuPathDB Galaxy workshop instance.**

If you do not have an account with VEuPathDB/FungiDB, please create one now.

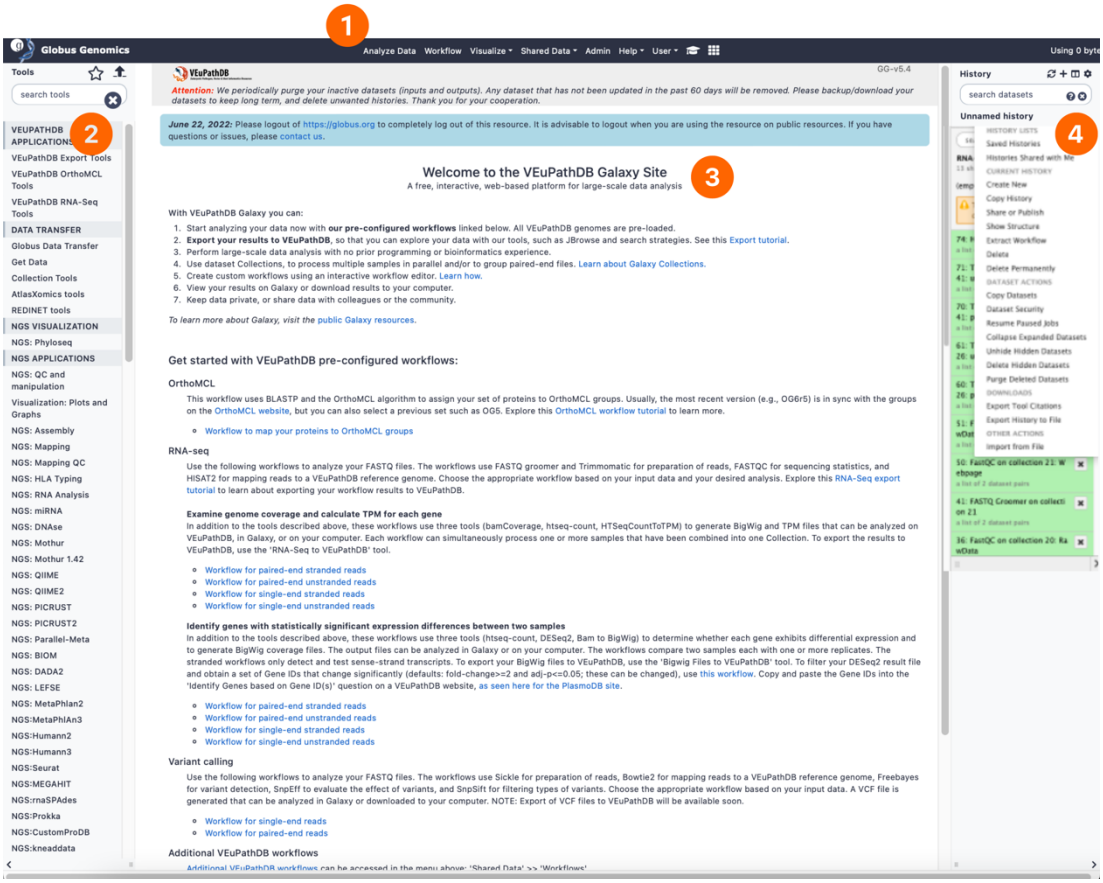
1. Click on the following URL to begin: https://veupathdb1.globusgenomics.org/
2. On the next page, you will be asked to define your organization. Choose the "VEuPathDB" option and click on the "Continue" button.
3. If you are not already logged into VEuPathDB, you will be prompted to do so.
4. Click on "Continue" on the next page (no need to link an existing account).
5. Select the "non-profit" option and agree to the Terms of Service. Click continue.
6. The next page will ask for permissions required to use this Galaxy instance. Click on "Allow"



The anatomy of the VEuPathDB Galaxy landing page.

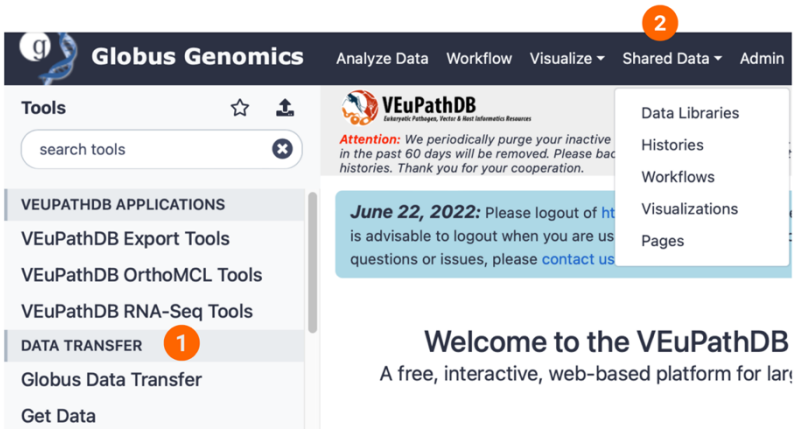
The workspace has four major components:

1. The top menu controls the main interface, provides access to the landing page, shared data, public and private workflows & more.
2. The left panel has a list of available tools where the VEuPathDB export tools are listed at the top.
3. The main welcome (landing) page is the interactive interface that houses pre-configured workflows, workflows editor, etc.
4. The panel on the right provides access to histories, deleted datasets, and other useful functions, including options to delete and purge datasets.



Don't see a tool you need for your research? – Let us know by sending an email to help@fungidb.org

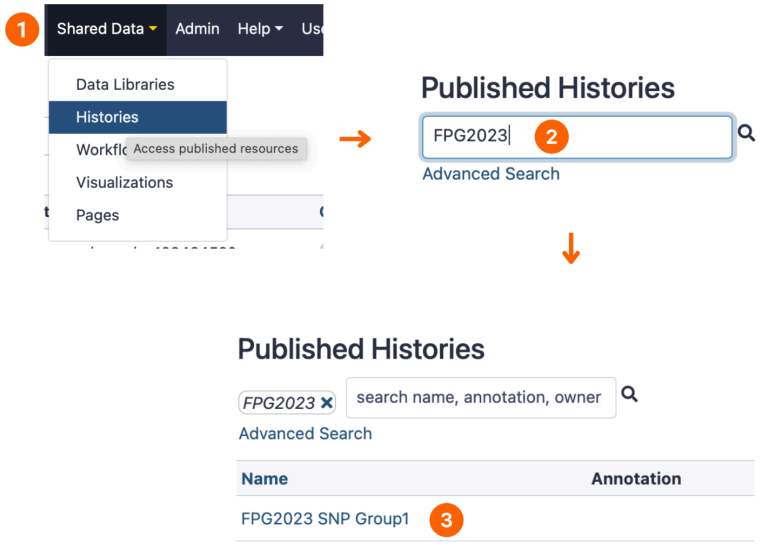
Importing data for your workflow.
There are multiple ways to important data into your Galaxy workspace. You can transfer data via tools located under the "Data Transfer" section in menu on the left (1). You can also transfer data from the "Shared Data" section in the main menu (2). The latter prides access to pre-loaded raw data, publicly shared workflows, or workflow results (histories), etc.



For this exercise, pre-loaded raw files should be imported from the "Shared Data" > Histories.
Only one person per each group should import data files and deploy an SNP workflow. Everyone will practice data analysis in NGS Part 2 module. For group assignments, see below.

- **Import data for your SNP workflow via the Shared histories option.**

1. From the top menu, select "Shared Data > Histories" option.
2. Filter all public workflows on "FPG2023" .
3. Click on the history link that correspond to your group number (e.g., FPG2023 SNP Group1) to import the data into your Galaxy workspace.



Group assignments

Groups 1 *Aspergillus fumigatus*. Paired-end data. A clinical isolate from pleural fluid of a patient. Isolate: AFIS2503.

History name for download (in Galaxy)	FPG2023 SNP Group1
Ref genome (in Galaxy)	FungiDB-29_AfumigatusAf293_Genome

Groups 2 *Aspergillus fumigatus*. Paired-end data. A clinical isolate from pleural fluid of a patient. Isolate: AFIS1415.

History name for download (in Galaxy)	FPG2023 SNP Group2
Ref genome (in Galaxy)	FungiDB-29_AfumigatusAf293_Genome

Group 3 *Zymoseptoria tritici*. Paired-end data. An isolate collected from common wheat (Triticum aestivum) in Switzerland: Eschikon. Isolate: ST16CH_1A27.

History name for download (in Galaxy)	FPG2023 SNP Group3
Ref genome (in Galaxy)	FungiDB-34_ZtriticilPO323_Genome

Group 4 *Zymoseptoria tritici*. Paired-end data. An isolate collected from common wheat (Triticum aestivum) in Oregon: USA. Isolate: ORE15_Mad_G1.

History name for download (in Galaxy)	FPG2023 SNP Group4
Ref genome (in Galaxy)	FungiDB-34_ZtriticilPO323_Genome

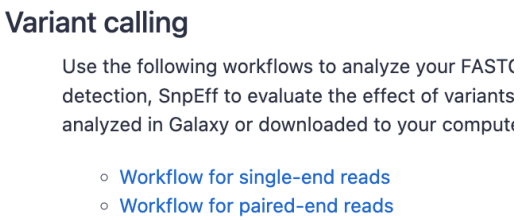
Group 5 *Candida auris*. Paired-end data. An isolated collected from an apple surface in India. Isolate: VPCI-F37-B-2021.

History name for download (in Galaxy)	FPG2023 SNP Group5
Ref genome (in Galaxy)	FungiDB-37_CaurisB8441_Genome

Group 6 *Candida auris*. Paired-end data. An isolated collected from an apple surface in India. Isolate: VPCI-F1-A-2020.

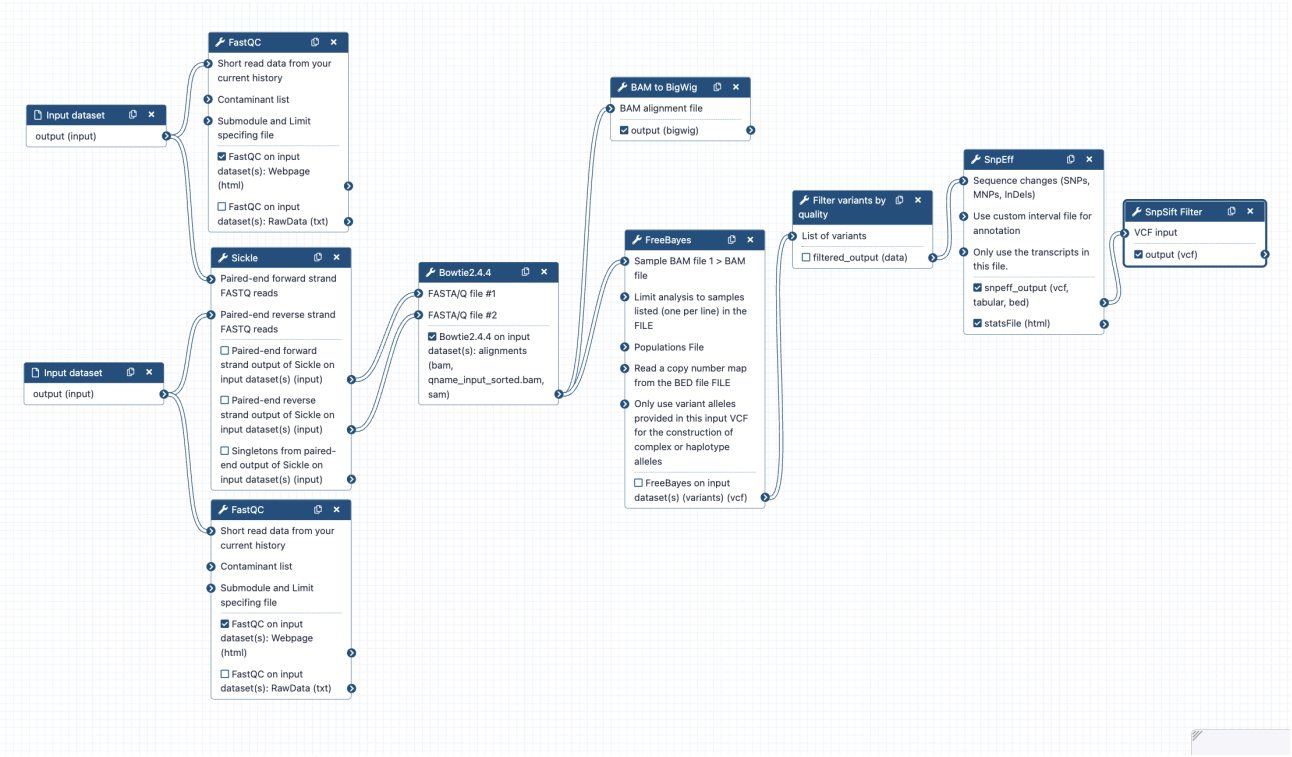
History name for download (in Galaxy)	FPG2023 SNP Group6
Ref genome (in Galaxy)	FungiDB-37_CaurisB8441_Genome

Once the data files have been transferred into your galaxy history you need to choose a workflow appropriate for your data (paired or single -read).



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.



- **Set workflow parameters.**
    1. For paired-end data, make sure that the input steps are set to the *xxxx_1.fastq.gz* and *xxxx_2.fastq.gz* as by default both have the same one selected. Here is an example (disregards "7" and "8" and it simply refers to the ordered file number).

Note: for single read data, you will have only one file.

2. Select the correct reference genome for Bowtie2 (see group assignment above).
3. Select the correct reference genome for FreeBayes (see group assignment above).
4. Select the correct reference genome for SnpEff (see group assignment above).
5. Click Run Workflow.