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.

RNA sequence data analysis via VEuPathDB Galaxy, Part I

Learning objectives:

- Become familiar with the VEuPathDB Galaxy workspace.
- Create collections of datasets from the pre-loaded data.
- Run a pre-configured RNA-Seq workflow.

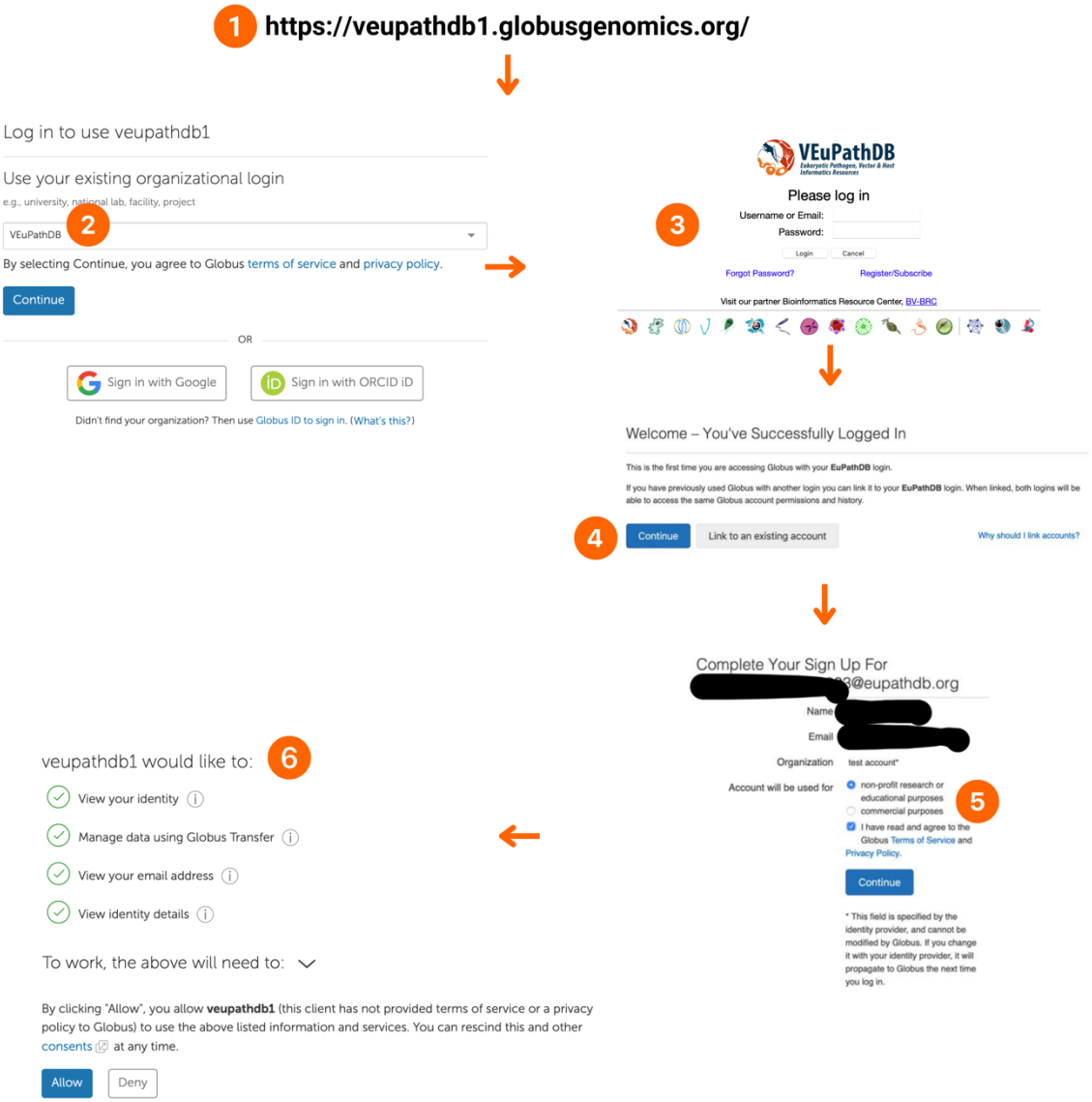
For this exercise, we will retrieve raw sequence files from the "shared history" section in VEuPathDB Galaxy and then run files through a pre-configured RNA-Seq workflow that will align the data to a reference genome, calculate expression values and determine differential expression.

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.

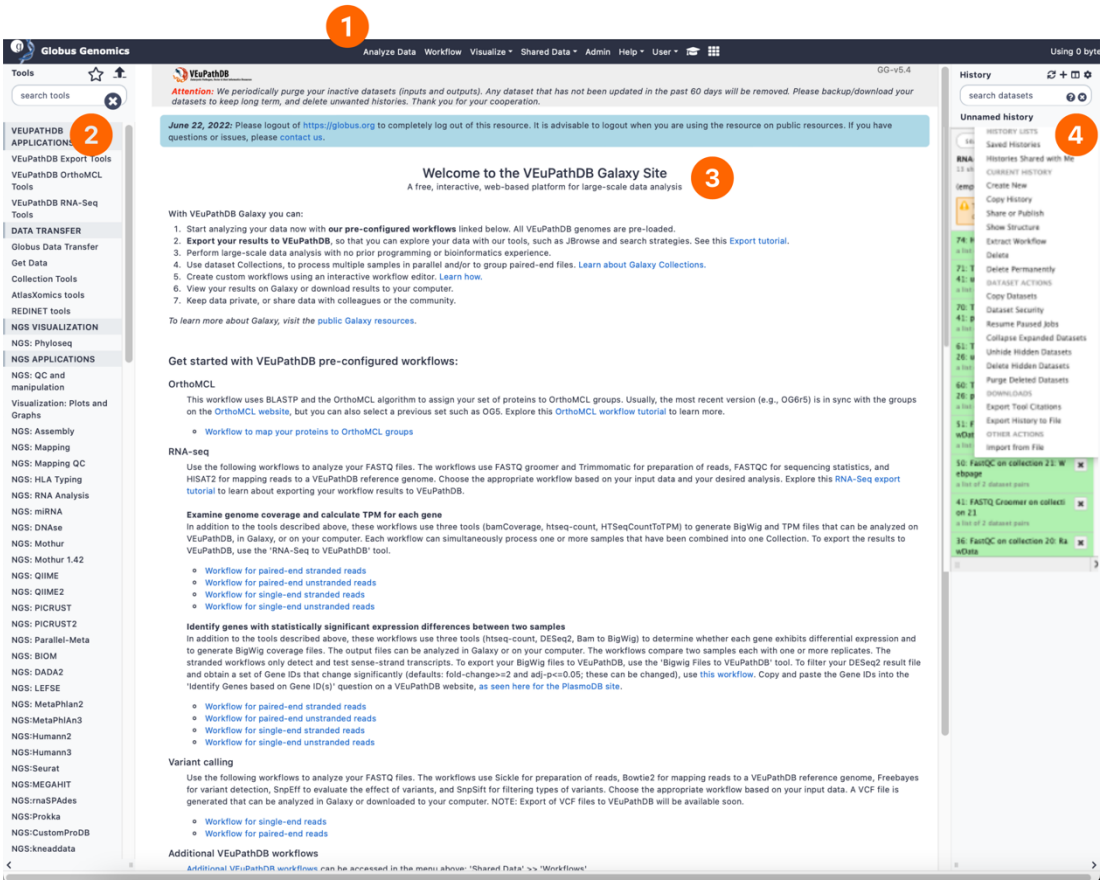
1. Click on the following URL to begin: https://veupathdb1.globusgenomics.org/
2. On the next page, you will be asked to define your organization. Choose the "VEuPathDB" option and click on the "Continue" button.
3. If you are not already logged into VEuPathDB, you will be prompted to do so.
4. Click 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:

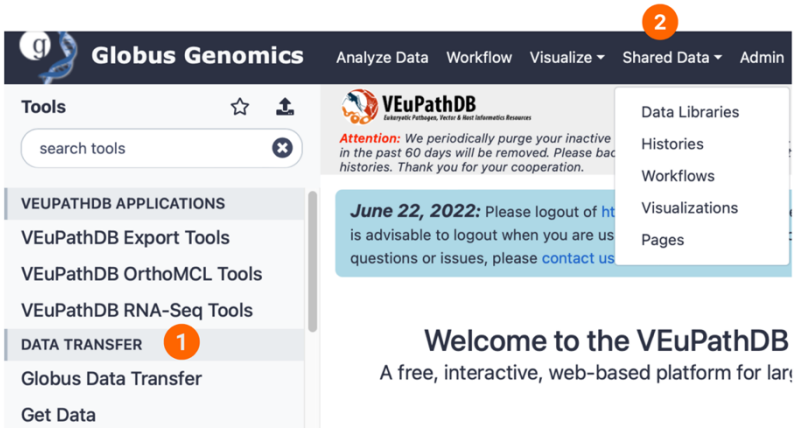
1. The top menu controls the main interface, provides access to the landing page, shared data, public and private workflows & more.
2. The left panel has a list of available tools where the VEuPathDB export tools are listed at the top.
3. The main welcome (landing) page is the interactive interface that 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.

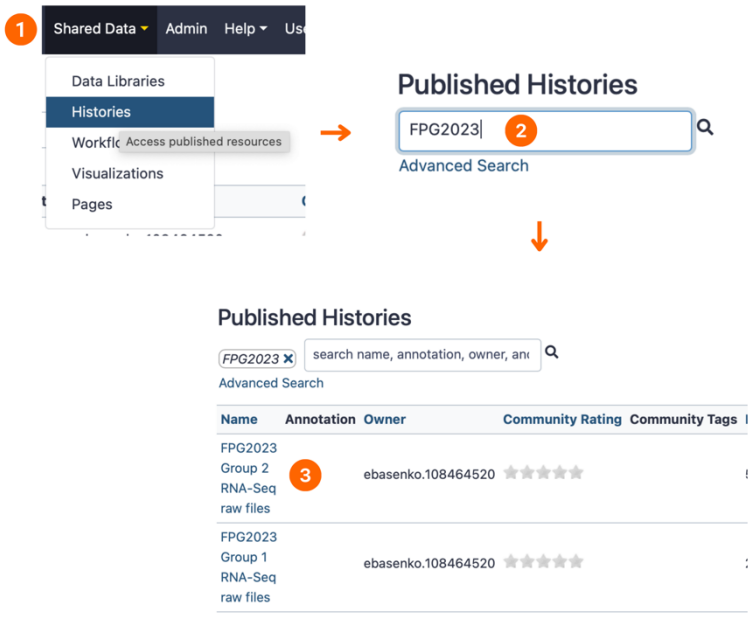
There are multiple ways to important data into your Galaxy workspace. You can transfer data via tools located under the "Data Transfer" section in menu on the left (1). You can also transfer data from the "Shared Data" section in the main menu (2). The latter prides access to pre-loaded raw data, publicly 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 RNA-Seq workflow. Everyone will practice data analysis in NGS Part 2 module. For group assignments, see below.

- **Import data for your RNA-Seq 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 to import the data into your Galaxy workspace.



Group assignments (see more information about the files below)

Groups 1 & 2 *Aspergillus fumigatus*. Paired-end data. Analyze transcriptomes from cells incubated in human blood (B) and defined minimal media (M) for 30 and 180 min.

Group Number	1	2
Comparison	M30 vs B30	B30 vs B180
History name for download (in Galaxy)	FPG2023 Group 1 RNA-Seq raw files	FPG2023 Group 2 RNA-Seq raw files
Ref genome (in Galaxy)	FungiDB-29_AfumigatusAf293_Genome	

Reference: PMID: 26311470 BioProject: PRJNA287921

Group 3 *Candida parapsilopsis.* Paired-end data. Analyze transcriptomes from cells grown under planktonic and biofilm-inducing conditions. Control: planktonic.

Comparison	Planktonic vs Biofilm
History name for download (in Galaxy)	FPG2023 Group3 RNA-Seq raw files
Ref genome (in Galaxy)	FungiDB-42_CparapsilosisCDC317_Genome

Reference: PMID: 25233198 BioProject: PRJNA246482

Group 4 *Coccidioides posadasii*. Single read data. Analyze transcriptomes from mycelia (non-pathogenic stage) and spherules (pathogenic stage).

Comparison	Mycelia vs Spherules
History name for download (in Galaxy)	FPG2023 Group 4 RNA-Seq raw files
Ref genome (in Galaxy)	FungiDB-61_CposadasiiSilveira2022_Genome

Reference: PMID: 22911737 BioProject: PRJNA169242

Group 5 *Fusarium graminearum*. Paired-end data. Analyze spore and mycelial transcriptomes.

Comparison	Spores vs Mycelia
History name for download (in Galaxy)	FPG2023 Group 5 RNA-Seq raw files
Ref genome (in Galaxy)	FungiDB-31_FgraminearumPH-1_Genome

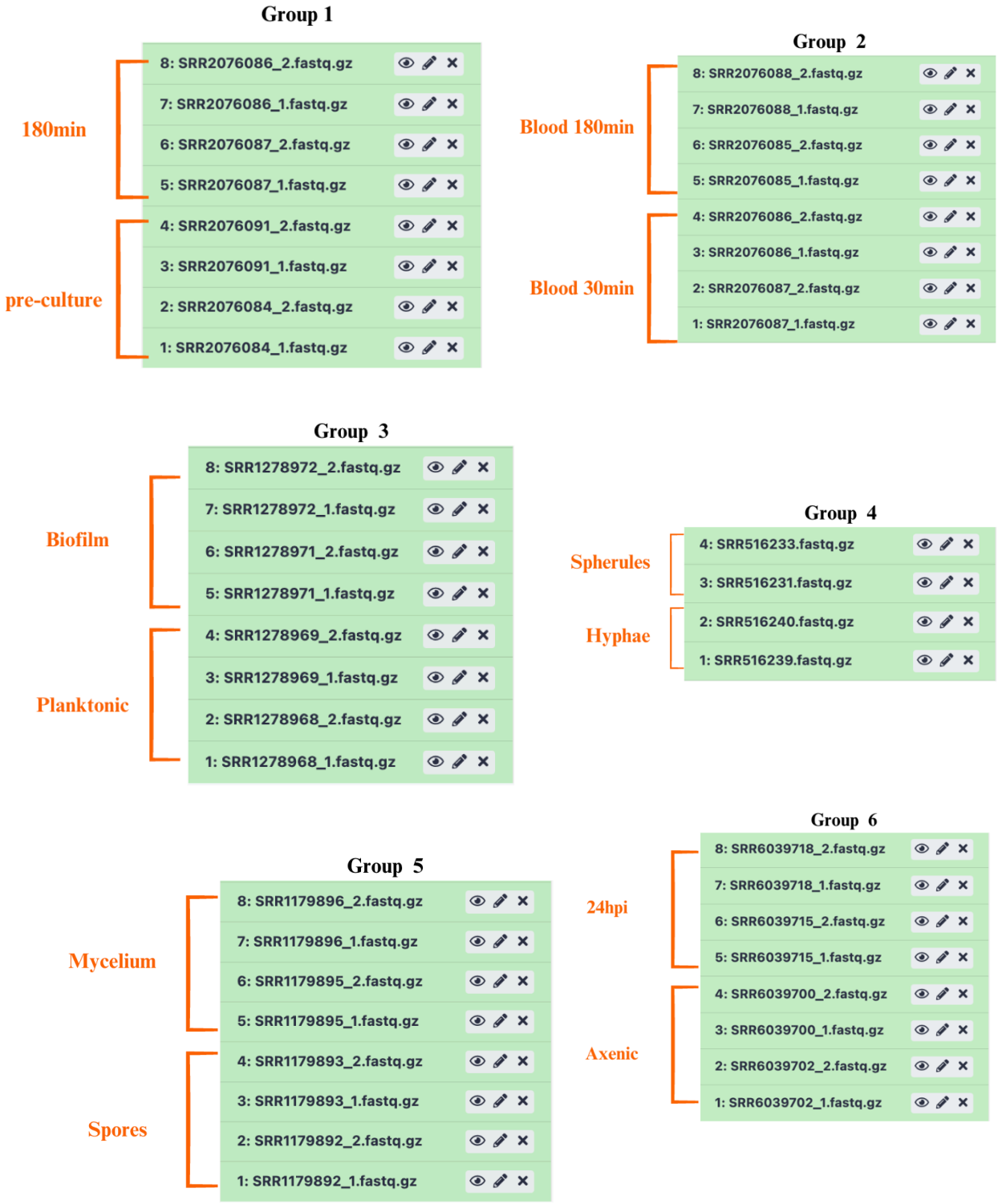
Reference: PMID: 24625133 BioProject: PRJNA239711

Group 6 *Ustilago maydis*. Paired-end data. Analyze transcriptomes from plant-associated development samples (axenic culture vs 12 days post infection (dpi)).

Comparison	0h vs 12 dpi
History name for download (in Galaxy)	FPG2023 Group 6 RNA-Seq raw files
Ref genome (in Galaxy)	FungiDB-51_Umaydis521_Genome

Reference: PMID: 33653886 BioProject: PRJNA407369

Guide to FPG2023 RNA-Seq histories and file organisation.



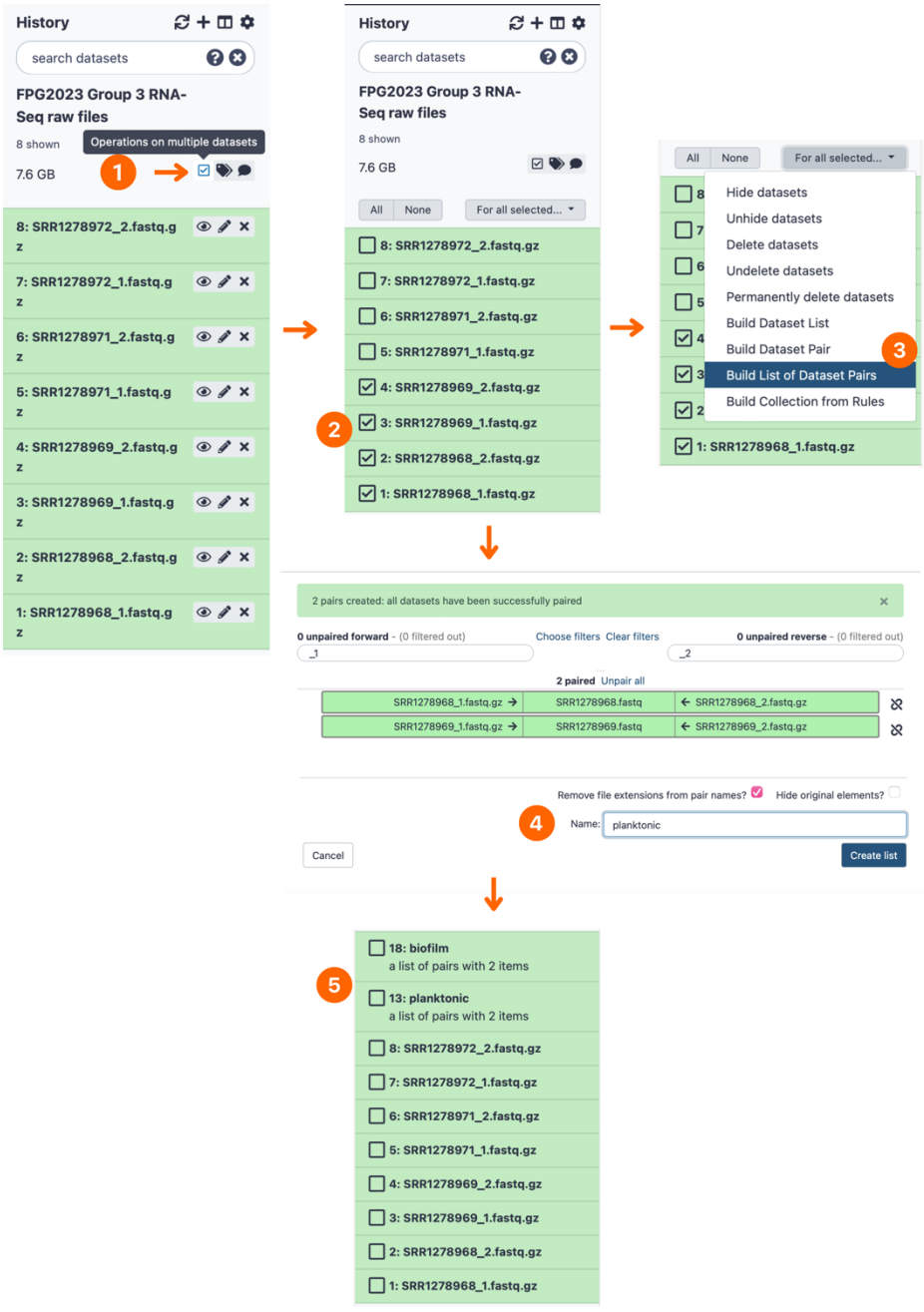
Each dataset contains two replicates. For datasets with multiple samples (e.g., containing biological replicates), it is useful to organize them into "Collections" (e.g., spore and mycelia). Organizing samples with replicates into collections also reduces the complexity Galaxy workflows.

- **Organize samples with replicates into collections:**

    1. Click on the checkbox function “operation on multiple datasets”.
    2. Select samples that belong to the same condition (control samples will appear at the bottom, see file mapping notes for each group below).
    3. Click on “For all selected” and choose “Build List of Dataset Pairs”.

    Note: for single read data, choose “Build Data List” option instead.

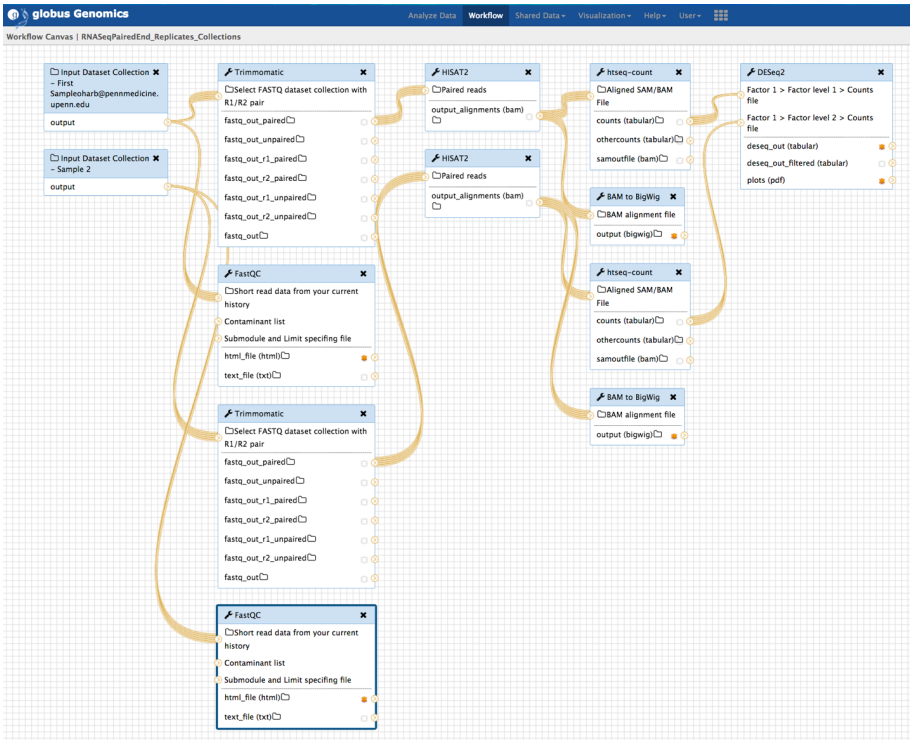
    4. Name the sample (e.g. planktonic) and click “Create List”. Note: Usually the correct pairs are auto selected.
    5. Repeat for the comparator sample. You should end up with 2 datasets (e.g., planktonic and biofilm).



Running a workflow in Galaxy

You can create your own workflows in galaxy using the tools from the menu on the left. For this exercise we will use a preconfigured workflow that consists of the following steps:

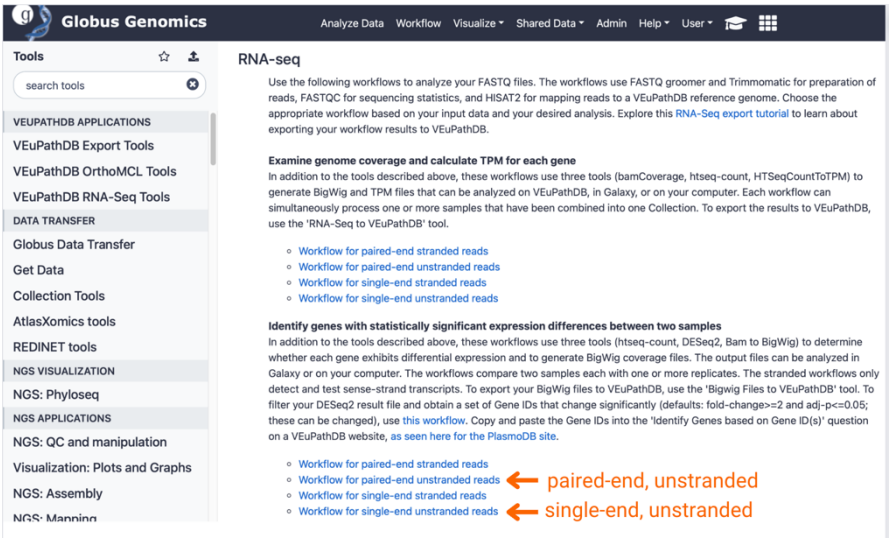
1. Input: raw data, dataset collections.
2. FASTQC: analyse for quality, generate read quality reports.
3. Trimmomatic: trims the reads based on their quality scores and adaptor sequences.
4. HISAT2: align reads to a reference and generate coverage plots.
5. HTSeq: estimate abundance (read counts per gene), generate coverage plots for JBrowse (BAM to BigWig).
6. DESeq2: differential expression of genes between samples.



- **Deploy a pre-configured workflow.**

To do this, navigate to the Galaxy home page and select the workflow appropriate for your dataset:

  - For paired-read datasets choose “Workflow for paired-end unstranded reads”.
  - For single read data, choose “Workflow for single-end unstranded reads”.

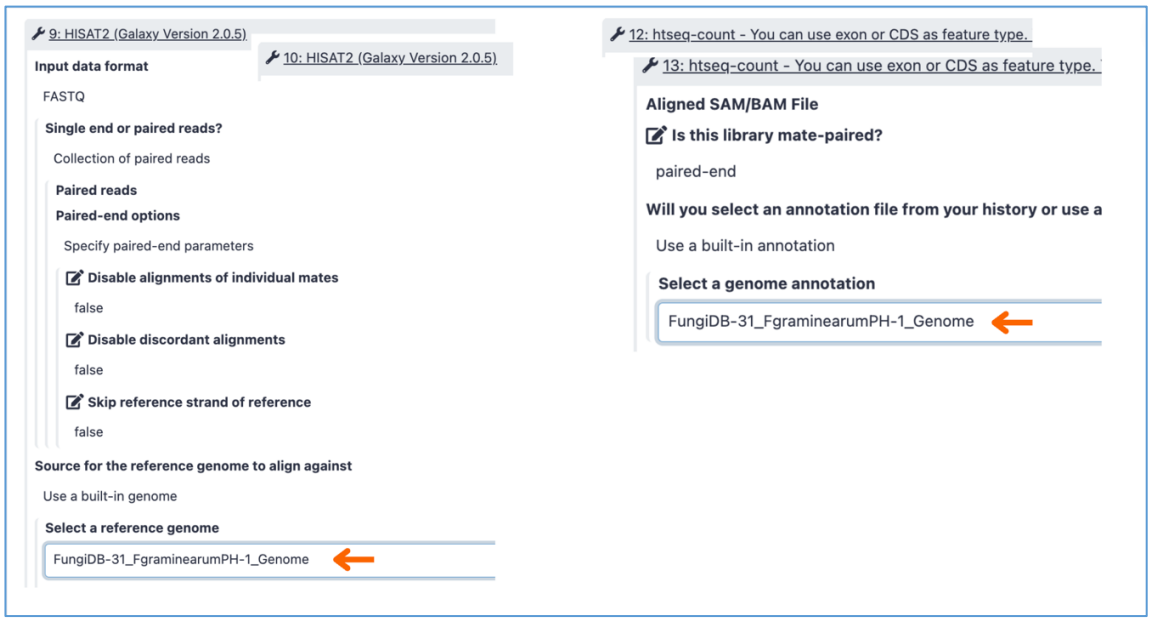


- **Configure an RNA-Seq workflow.**

There are multiple steps in the workflow, but you do not need to configure all of them. For this exercise, you will need to configure the following:

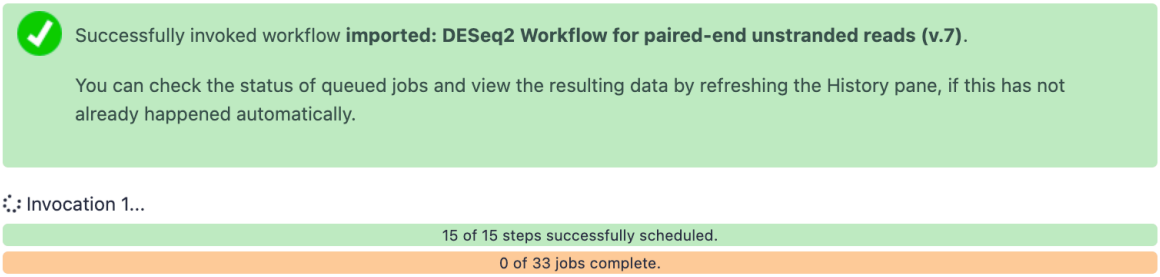
1. Input dataset collection 1 (e.g., planktonic).
2. Input dataset collection 2 (e.g., biofilm).
3. Both HISAT2 steps (requires reference genome – refer to the group assignments section above for this info).
4. Both htseq-count steps (requires reference genome – refer to the group assignments section above for this info).
5. DESeq2 (requires reference genome – refer to the group assignments section above for this info).

Make sure to set the correct reference genomes for HISAT2, htseq-count, and DESeq2 steps. It is critical that you select the correct genome that matches the experimental organism for your samples:



Name you factor levels. This helps keep everything organized and name properly in the workflow. Each factor level is typically the name of the condition, like “mycelia” or “spore”.

a. Once you are sure everything is configured correctly, click on “Run Workflow” at the top.



The steps will start running in the history section on the right. Grey means they are waiting to start. Yellow means they are running. Green means they have completed. Red means there was an error in the step.

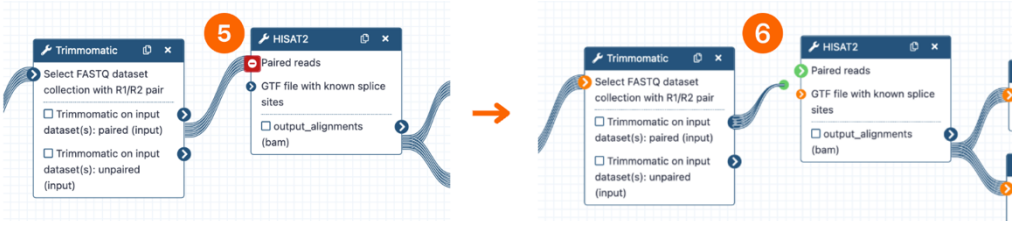
How to work with Galaxy editor (optional)

You can create your own workflows. The tools can all be added and configured in an interactive workflow editor.

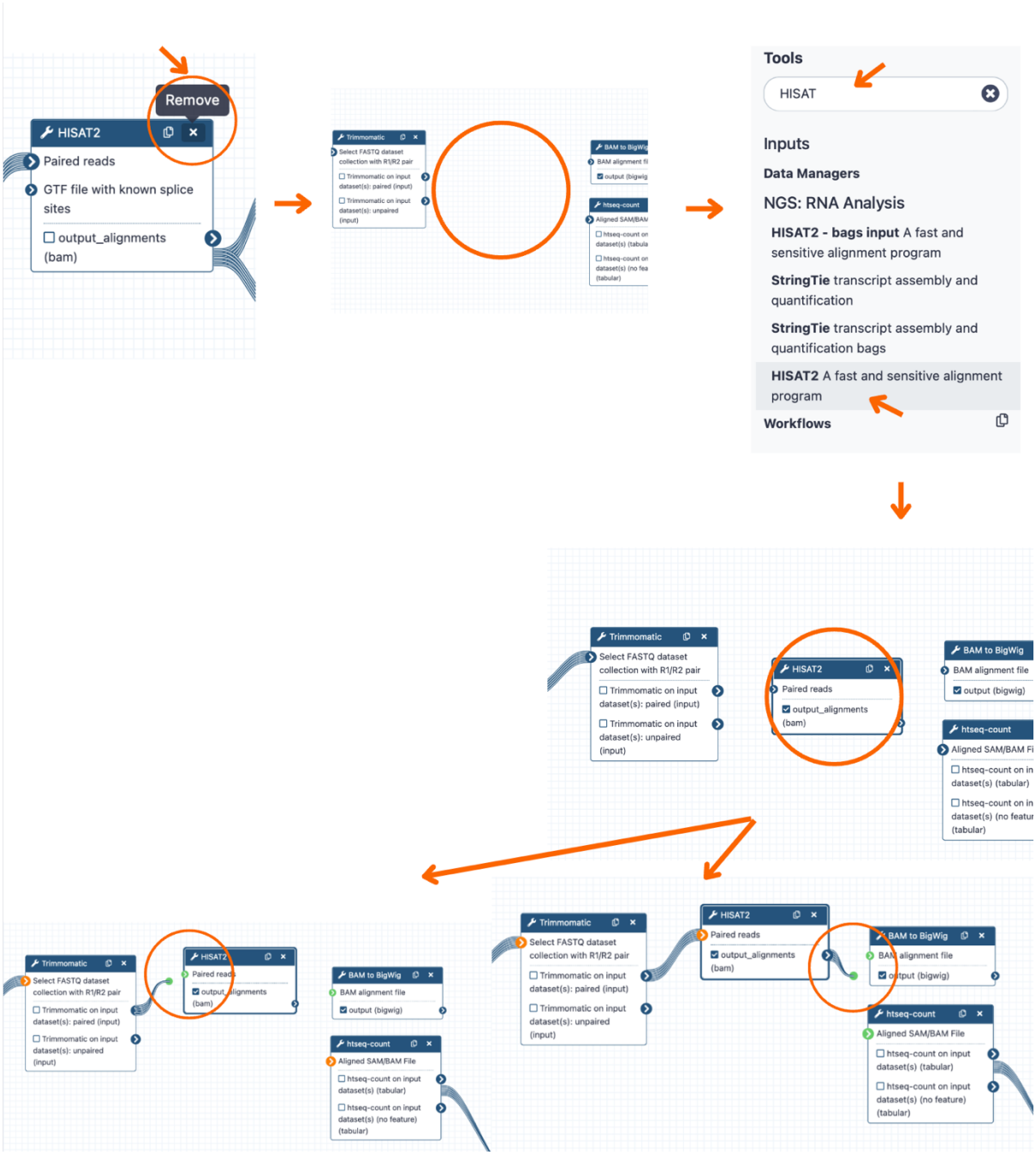
1. Navigate to the “Shared Data” menu.
2. Click on the “Workflows”.
3. Left-click on the “FPG2023 workflow editor practice” work to “import”
4. Once the workflow is important into your workspace, left-click and select “edit.

Once you are in the workflow editor:

5. Delete the Trimmomatic - HISAT2 connection.
6. Re-establish the connection by linking the “Trimmomatic on input dataset(s): paired (input) step to the “Paired reads” option in the HISTAS2.



7. Delete HISAT2 step completely by clicking on the “x” in the top right corner and use the tools menu on the left to insert it back.



Note: Sometimes you may be unable to re-establish connection. When this happens, take a look at the tool documentation notes in the right panel, check you r selection for single-read or paired-end setting in particular (paired-end setting must be selected if you are dealing with reverse and forward reads).

Now that you have learned the principals of workflow editing, you can either practice saving the workflow by clicking on the wheel at the far top corner or simply existing the workflow editor without saving.