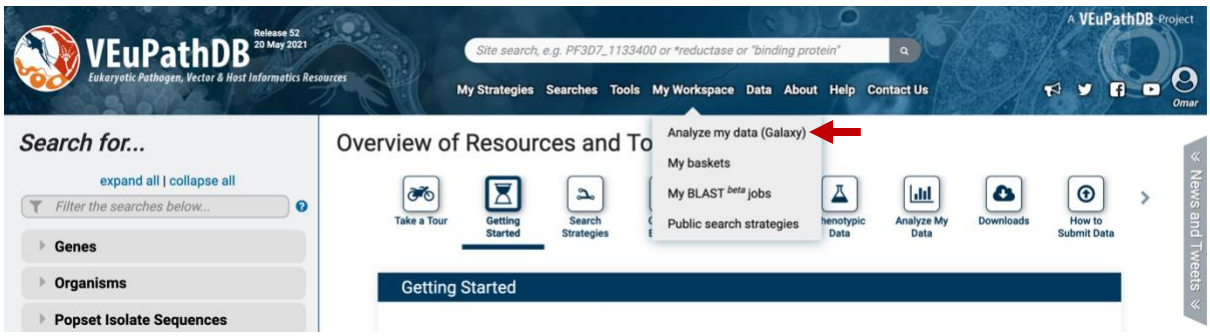
RNA sequence data analysis via VEuPathDB Galaxy, Part I
Uploading data and starting the workflow (Group Exercise)

Learning objectives:

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

VEuPathDB Galaxy-based workspace offers pre-loaded genomes, private data analysis and display, and the ability to share and export analysis results and also import certain datasets into private workspace within VEuPathDB (My Datasets section).

VEuPathDB Galaxy workspace can be accessed from the *My Workspace* tab on the home page of FungiDB or any other VEuPathDB site. To log in, users must have an account with FungiDB/VEuPathDB, which is free. After an account is created, users receive access



to the VEuPathDB Galaxy services and tools.

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

Galaxy is an open, web-based platform for data intensive biomedical research. Galaxy allows you to perform, reproduce, and share complete analyses without the use of command line scripting. VEuPathDB developed its own Galaxy instance in collaboration with Globus Genomics. Many resources are available to learn how to use Galaxy. The following link has information about additional resources to help you learn how to use Galaxy:

https://wiki.galaxyproject.org/Learn#Galaxy_101

For this exercise, we will retrieve raw sequence files from a repository, assess the quality of the data, and then run the data through a workflow (or pipeline) that will align the data to a reference, calculate expression values and determine differential expression.

Part 1, uploading data and starting the workflow will be performed today. The workflows will run overnight and we will view / interpret the results tomorrow in Part 2.

We will be working in groups. Each group will have 2-3 members. One person in the group will run the Galaxy controls on one computer. The other members' roles are to ensure that the correct datasets are used and that the correct workflow parameters are selected.

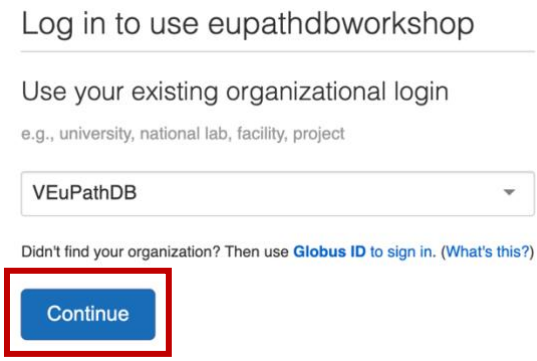
Section I: Setting up your VEuPathDB Galaxy account

Step 1: Access the VEuPathDB Galaxy instance at the following URL:

Use the link below only for the workshop – this is a special instance for our training

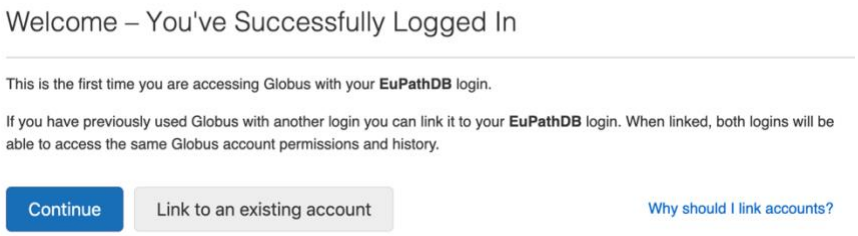
https://veupathdb1.globusgenomics.org/

Step 2: On the next page you will be asked to define your organization. Choose VEuPathDB and click Continue.

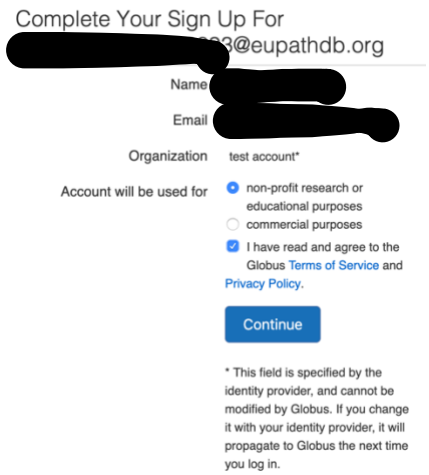


Step 3: If you are not already logged into VEuPathDB you will be prompted to do so now.

Step 4: Click on “continue” on the next page (no need to link an existing account).

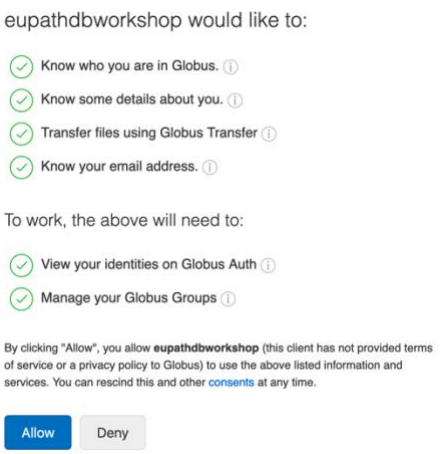


Step 5: on the next window select the “non-profit” option and agree to the Terms of Service. Click continue.



Step 6: The next page will ask for permissions required to use this Galaxy instance. Click on “Allow”

Step 5: Congratulations, you are in!



The anatomy of the VEuPathDB Galaxy landing page.

The workspace has four major components:

a) the top menu controls the main interface
b) the left panel has a list of available tools
c) the main welcome page is the interactive interface that houses pre-configured workflows, workflows editor, etc.
d) the right panel provides access to histories, deleted datasets, and other useful functions

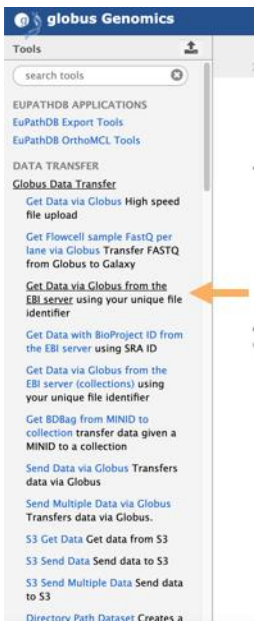
Section II: Importing data to Galaxy

There are multiple ways to important data into your Galaxy workspace. For this exercise, we will use the '**Get Data via Globus from the EBI: server using your unique file identifier**"
tool and enter the sequence repository sample IDs based on your group assignments (below). *Remember only one person in your group will be running the workflow*. Although all group members can sign up for an account for later use, please only one person should start a workflow today because we do not want to overload the servers. The samples below were all generated by paired end sequencing; hence each sample ID will result in transferring two files to your galaxy history. The files are fastq files that are compressed (that is why they end in .gz = gzip).

Group assignments:

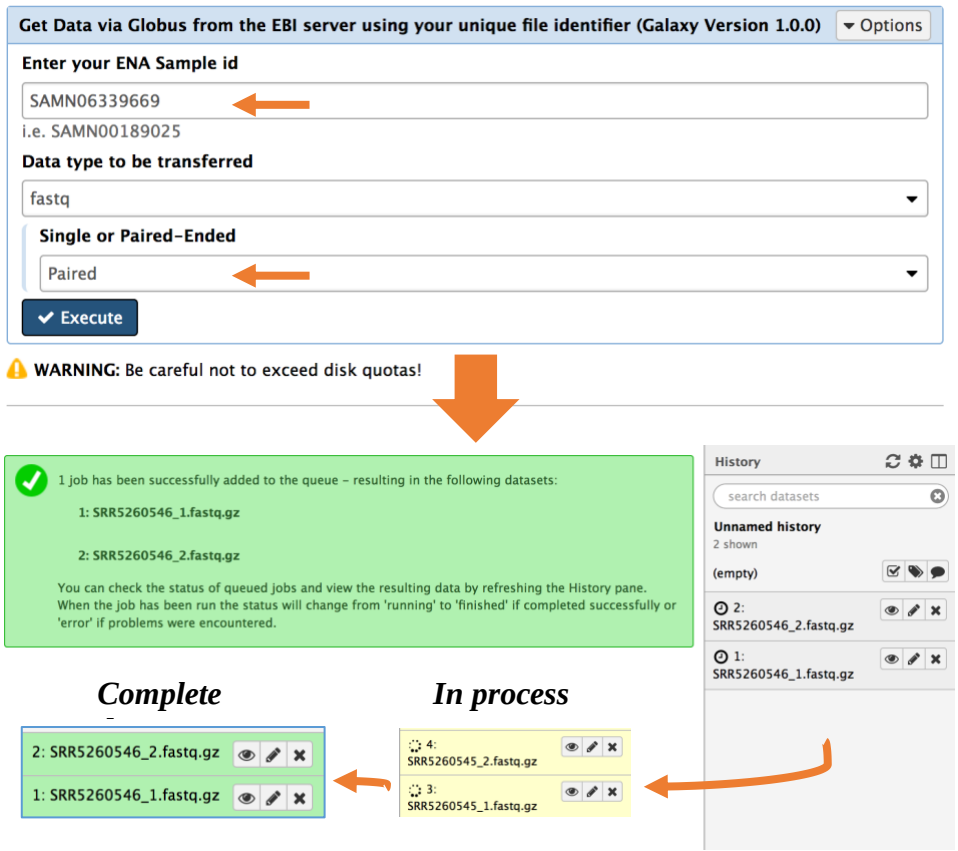
See separate group assignment sheet

Step 1: Click on the "**Globus Data Transfer**" link in the left-hand menu. This will reveal a list of options; click on "**Get Data via Globus from the EBI server**". ***important: do not select the option for transferring a collection.

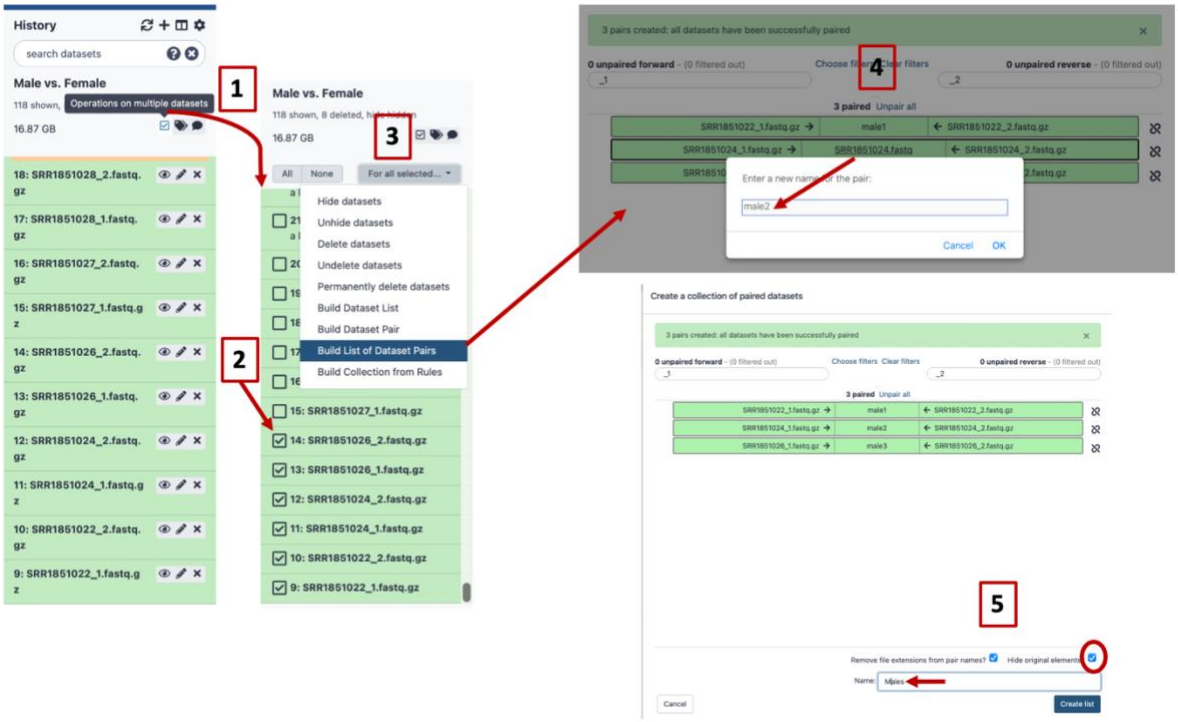


Step 2: In the middle section enter the sample ID and choose whether the run was single or paired end. Click on Execute.

Note that the sample ID resulted in importing two files one for each pair. Repeat this process for each sample you want to import. *If you are working with samples from two conditions and the experiment was done in triplicate and paired end sequenced then you should end up with 12 files; six from each condition.*



Step 3: If you are working with a dataset with biological replicates it is useful to organize the different conditions of your experiment into "Collections". For example, if your experiment included RNAseq from *Anopheles stephensi* males (three biological replicates) and females (three biological replicates), it is useful to organize these into two collections, one that includes all male insect files and the other that includes all the female files. Using collections also reduces the complexity of the Galaxy workflows. See below:

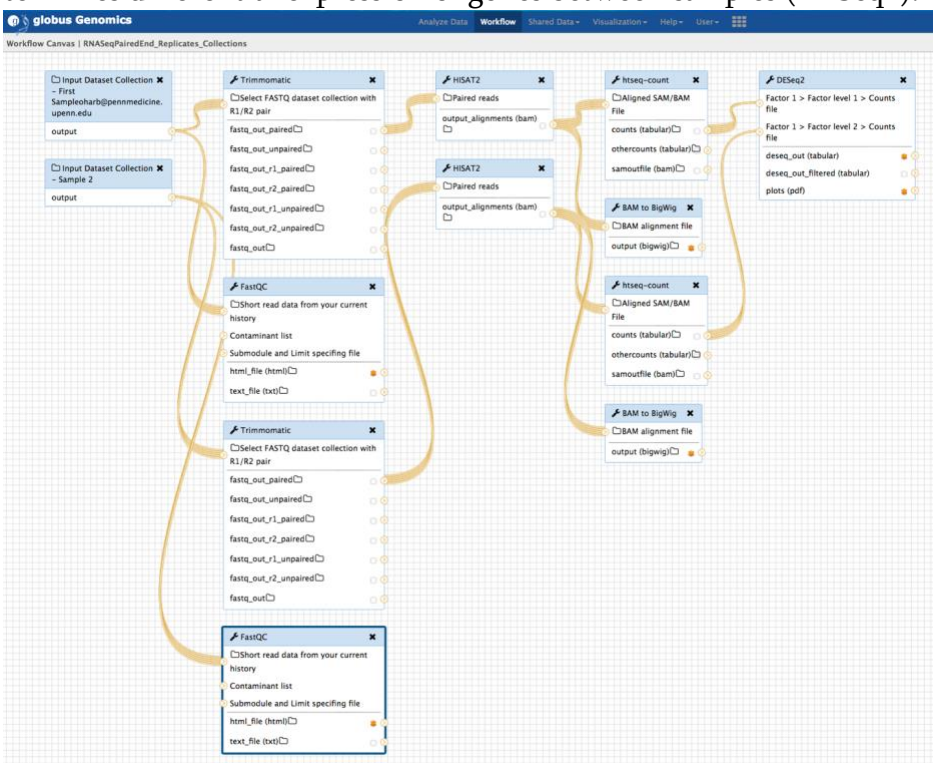


*Steps to create a collection. **1**. Click on the checkbox function "operation on multiple datasets". **2**. select fastq files that belong to the same condition. **3**.Click on the "For all selected" option and select "Build a list of dataset pairs". **4**. Give each pair a name by clicking on the ID in the middle then renaming it (eg. male1, male2, male3 etc…). **5**. Give the collection a name and select the checkbox "hide original elements", then click on create list.*

Section II: Running a workflow in Galaxy

You can create your own workflows in galaxy based on your needs. The tools in the left section can all be added and configured as steps in a workflow that can be run on appropriate datasets. For this exercise we will use a preconfigured workflow that does the following main things:

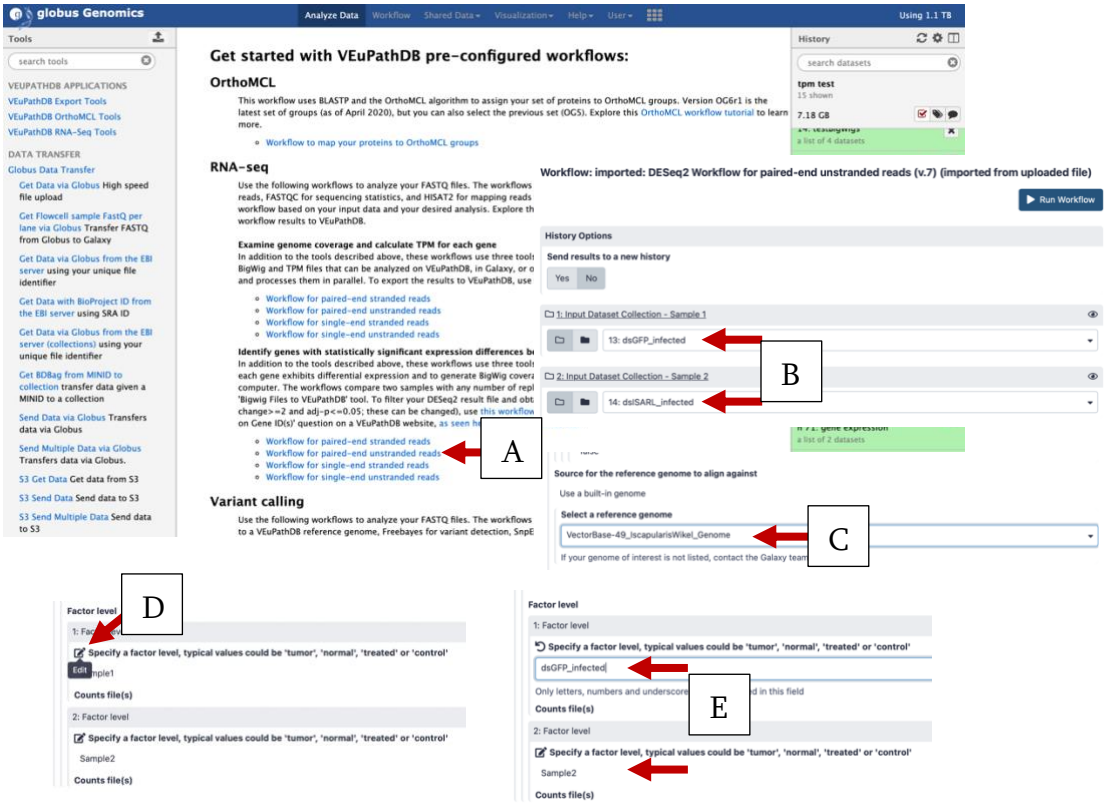
1. Analyzes the reads in your files and generates FASTQC reports.
2. Trims the reads based on their quality scores and adaptor sequences (Trimmomatic).
3. Aligns the reads to a reference genome using HISAT2 and generates coverage plots.
4. Determines read counts per gene (HTSeq)
5. Determines differential expression of genes between samples (DESeq2).



Additional resources:

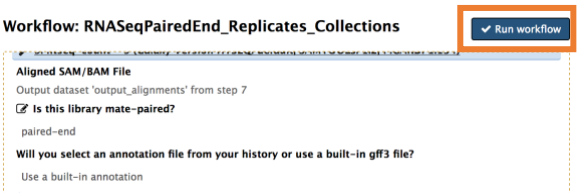
Galaxy Project (https://usegalaxy.org/)
Trimmomatic manual
FastQC
HISAT2
HTseq
DEseq2

To use one of the VEuPathDB preconfigured workflows, go to the Galaxy home page and select the workflow that you would like to run. For this exercise "**Workflow for paired-end unstranded reads**" – click on this workflow to run it (See A in figure below).



- Configure your workflow – there are multiple steps in the workflow, but you do not need to configure all of them. For the purpose of this exercise, you will need to configure the following:
    - Select the input dataset collections. These are the collections of fastq files you just created. Workflow steps 1-2 allow you to select the datasets. (**B above figure**)
    - Some tools in the workflow require that you select the reference genome to be used. In this workflow, both HISAT2 and HTSeq require this (note that each of these tools is in the workflow twice since you have two collections). It is critical that you select the correct genome that matches the experimental organism. So, for example, if your experiment was performed using *Plasmodium berghei i*ANKA, the reference genome you select should be *PlasmoDB-51_PbergheiANKA_Genome* (**C above figure**).
    - Name your factor levels. This helps keep everything organized and named properly in your workflow. Each factor level is typically the name of the condition, like Male or Female OR Susceptible or resistant (**D & E above figure**).

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



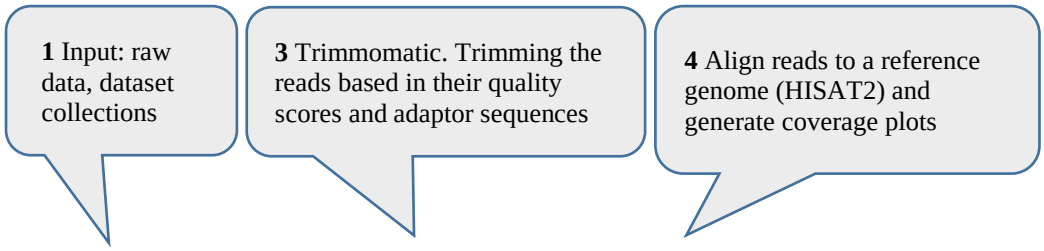
The steps will start running in the history section on the right. Grey means they are waiting to start. Yellow means they are running. Green means they have completed. Red means there was an error in the step.

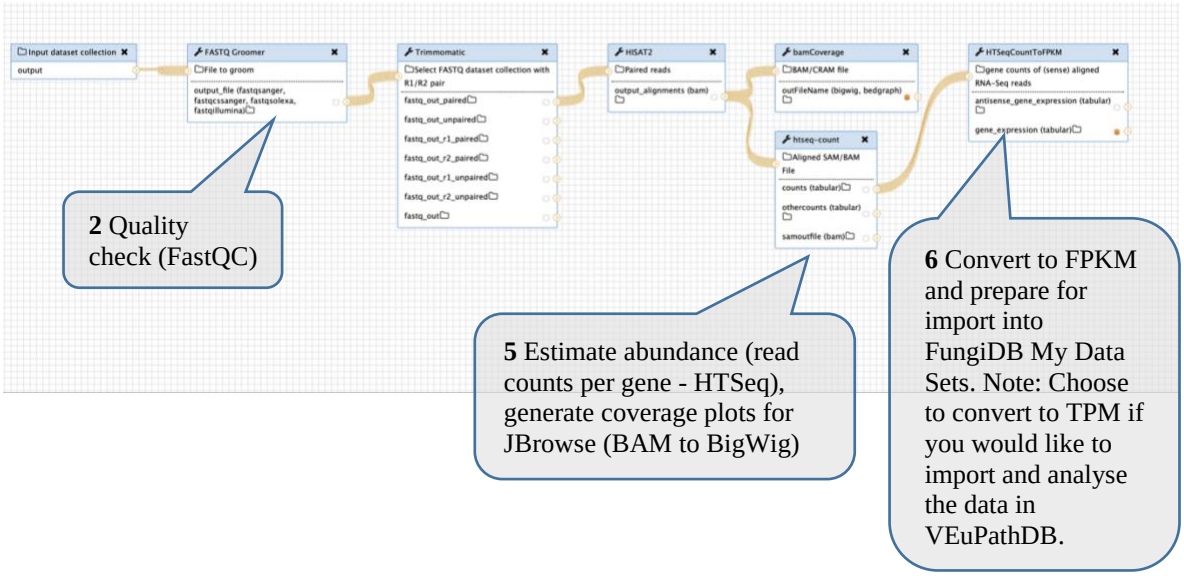
Practice working with Galaxy editor (optional)

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

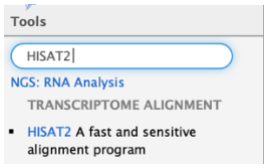
- Navigate to the Workflow tab from the main menu at the top and select
- Left click on the drop-down icon within the workflow you want to modify and select the “Edit” option.

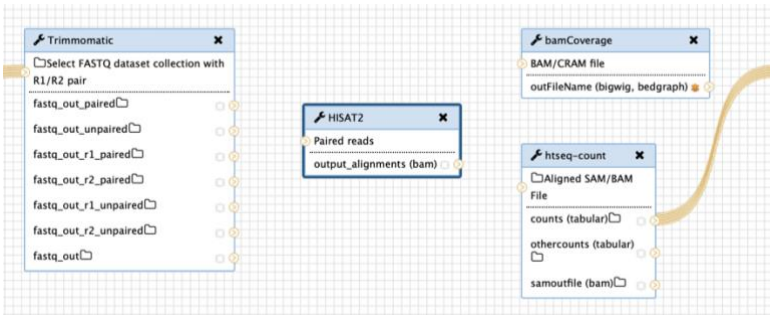
Sample workflow steps:





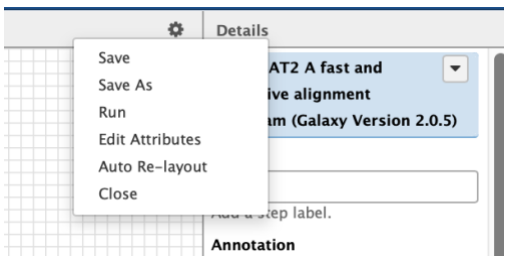
- Delete HISAT2 step by clicking on the " x " in the top right corner.
- Locate the HISAT2 tool in the Tools panel and click to insert it back into the workflow.
- Re-establish connections for HISAT2
    - Click on the arrow in the step before HISAT2 and drag to the appropriate input in HISAT2 tool.





- What happens? Can you reconnect it?

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.