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 any VEuPathDB site. To log in, users must have an account with VEuPathDB, which is free. After an account is created, users receive access to the VEuPathDB Galaxy services and tools.
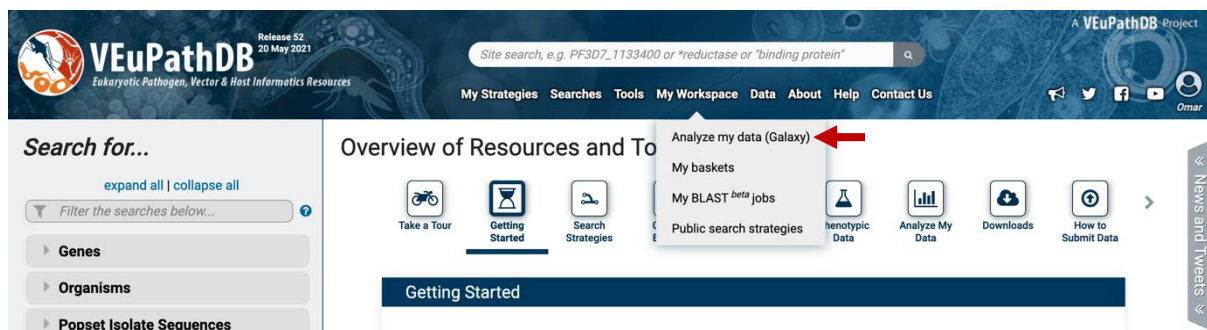
The Galaxy instance is not meant for long-term data storage. Datasets are automatically deleted after 60 days or when the total quota for all projects is reached. To save your data, download your analysis results locally and then *delete and purge* files to free up space for your next analysis.

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. One person in each 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.

Log in to use eupathdbworkshop

Use your existing organizational login
e.g., university, national lab, facility, project

VEuPathDB ▾

Didn't find your organization? Then use **Globus ID** to sign in. (What's this?)

Continue

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

VEuPathDB
Eukaryotic Pathogen, Vector & Host Informatics Resources

Please log in
Email:
Password:

Login   Cancel

Forgot Password?   Register/Subscribe

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

Welcome – You've Successfully Logged In

This is the first time you are accessing Globus with your **EuPathDB** login.

If you have previously used Globus with another login you can link it to your **EuPathDB** login. When linked, both logins will be able to access the same Globus account permissions and history.

Continue   Link to an existing account   Why should I link accounts?

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

Complete Your Sign Up For
apicoplast.250780093@eupathdb.org

Name	Omar harb*
Email	apicoplast@gmail.com*
Organization	test account*
Account will be used for	⦿ non-profit research or educational purposes
	○ commercial purposes
	☑ I have read and agree to the Globus Terms of Service and Privacy Policy.

Continue

* This field is specified by the identity provider, and cannot be modified by Globus. If you change it with your identity provider, it will propagate to Globus the next time you log in.

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

Step 5: Congratulations, you are in!

eupathdbworkshop would like to:

✓ Know who you are in Globus. ⓘ
✓ Know some details about you. ⓘ
✓ Transfer files using Globus Transfer ⓘ
✓ Know your email address. ⓘ

To work, the above will need to:

✓ View your identities on Globus Auth ⓘ
✓ Manage your Globus Groups ⓘ

By clicking "Allow", you allow **eupathdbworkshop** (this client has not provided terms of service or a privacy policy to Globus) to use the above listed information and services. You can rescind this and other consents at any time.

Allow   Deny

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

The menu at the top accesses the landing page, public and private workflows & more.

Main landing page with pre-configured VEuPathDB workflows that also serves as an interactive interface for creating and deploying workflows.

3

VEuPathDB Applications: Tools to data export into VEuPathDB sites

Sample workflows section

Tools (left panel): Section featuring all available tools. Don't see a tool? Let us know by sending an email to help@veupathdb.org

The history section provides access to workflow history, and much more, including options to delete and purge datasets

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 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.
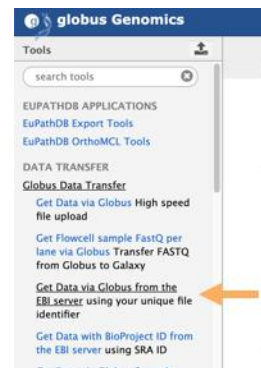
Step 2: In the middle section enter the sample ID and choose whether the run was single or paired end. Click on Execute. Note that the sample ID resulted in importing two files one for each pair. Repeat this process for each sample you want to import. *If you are working with samples from two conditions 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 workflow results. See below:

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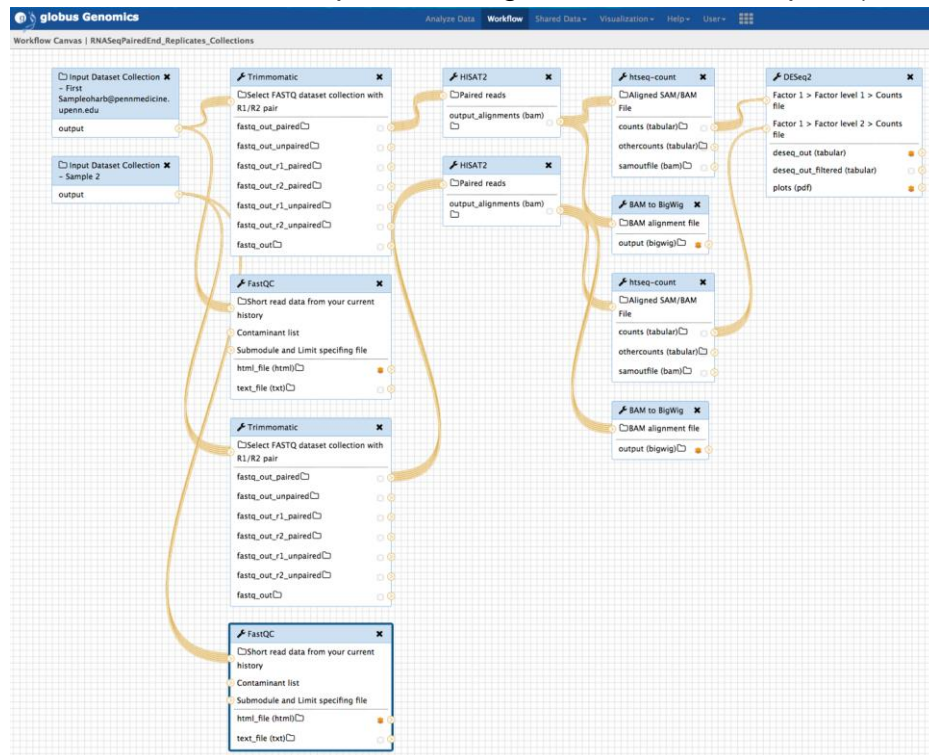
 Additional resources:
Galaxy Project (https://usegalaxy.org/)
Trimmomatic manual
FastQC
HISAT2
HTseq
DEseq2

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, we are using data with biological replicates which as suitable for statistical analysis. Choose the "**Workflow for paired-end unstranded reads**" under "**Identify genes with statistically significant expression differences between two samples**". (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**)
  - Select the reference genome for the alignments. 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**).

- o Name your factor levels for the DESeq2 statistical analysis. This helps keep everything organized and named properly in your workflow. Each factor level is typically the name of the condition or sample, like Male or Female OR Susceptible or resistant (**D & E above figure**).

- Once you are sure everything is configured correctly, scroll back up to the top and click "Run Workflow".

Workflow: RNASeqPairedEnd_Replicates_Collections   ✔ Run workflow

Aligned SAM/BAM File

Output dataset 'output_alignments' from step 7

☑ **Is this library mate-paired?**

paired-end

Will you select an annotation file from your history or use a built-in gff3 file?

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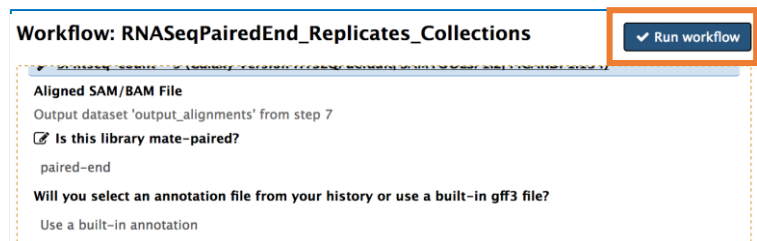
The steps will start running in the history section on the right. Grey means they are waiting to start. Yellow means they are running. Green means they have completed successfully. Red means there was an error in the step.
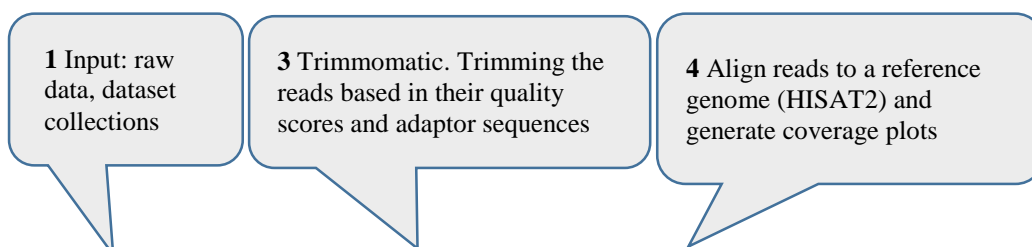

Practice working with Galaxy editor (optional)

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:

1 Input: raw data, dataset collections

3 Trimmomatic. Trimming the reads based in their quality scores and adaptor sequences

4 Align reads to a reference genome (HISAT2) and generate coverage plots

**2** Quality check (FastQC)

5 Estimate abundance (read counts per gene - HTSeq), generate coverage plots for JBrowse (BAM to BigWig)

6 Convert to FPKM and prepare for import into FungiDB My Data Sets. Note: Choose to convert to TPM if you would like to import and analyse the data in VEuPathDB.
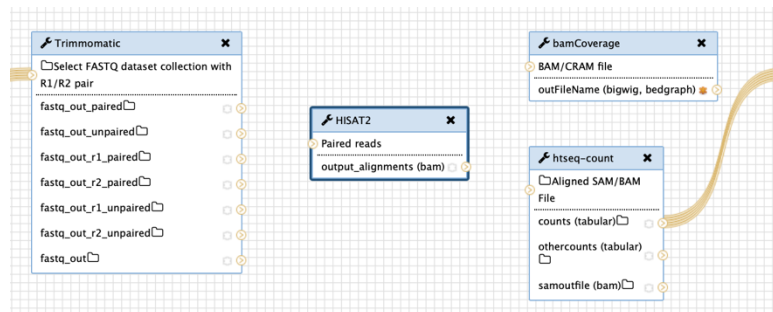
- Delete HISAT2 step by clicking on the " x " in the top right corner in the workflow.

- Locate the HISAT2 tool in the Tools panel and click to insert it back into the workflow.
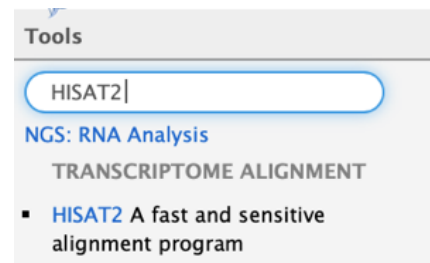

- Re-establish connections for HISAT2. Click on the arrow in the step before HISAT2 and drag to the appropriate input in HISAT2 tool.

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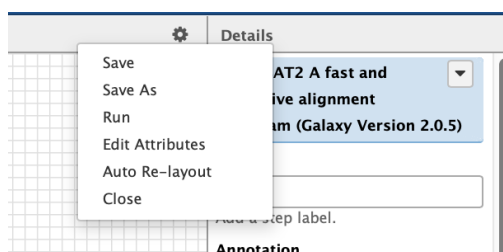
Now that you have learned the principals of workflow editing, you can either practice saving the workflow by clicking on the wheel at the far top corner or simply existing the workflow editor without saving.