RNA sequence data analysis via VectorBase Galaxy

Part I: Uploading data and starting the workflow

Learning objectives

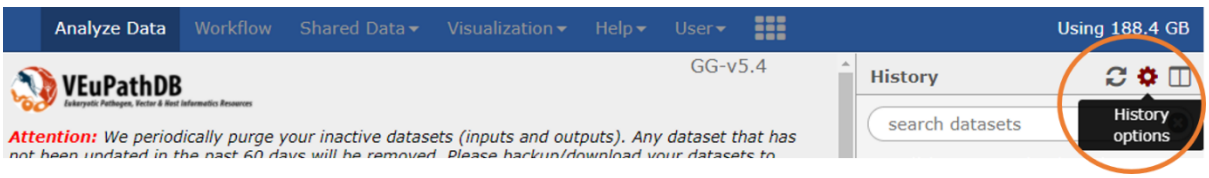
- Become familiar with VectorBase Galaxy workspace
- Import data from EBI to the VectorBase Galaxy
- Create collections of datasets
- Run a pre-configured RNA-Seq workflow

The VectorBase 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 VectorBase (*My Workspace* > My data sets).

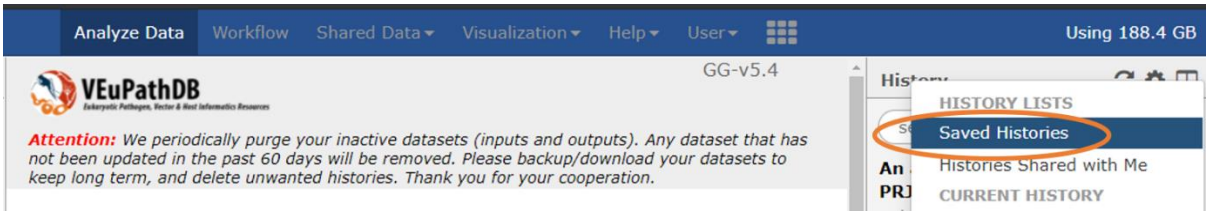
VectorBase Galaxy workspace can be accessed from the *My Workspace* tab or from the *Tools* menu tab, on the home page of VectorBase or any other VEuPathDB site. To log in, users must have an account with VectorBase/VEuPathDB, which is open access. After an account is created, users receive access to the VectorBase 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; to do it, follow these steps.

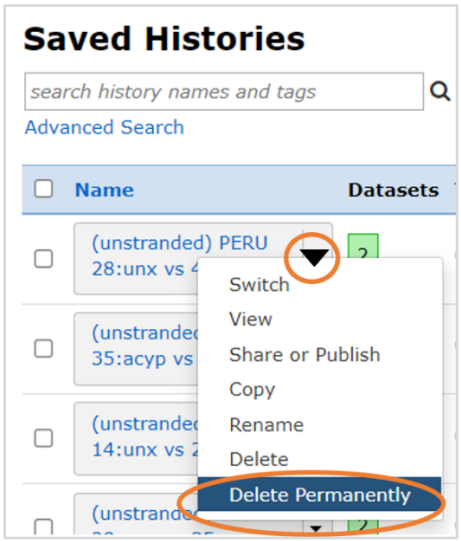
Gear icon > History options



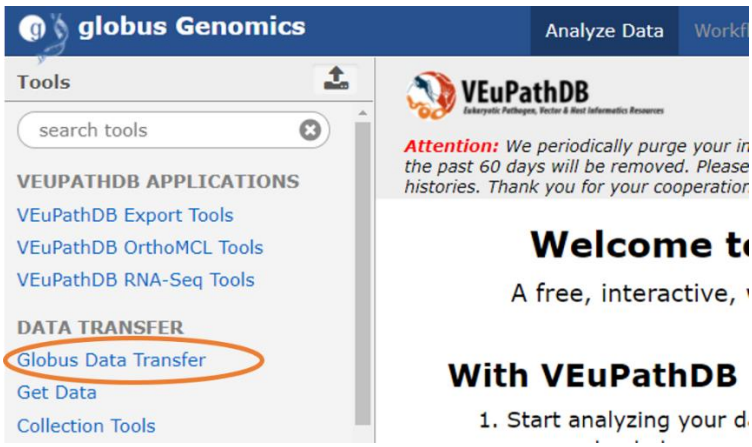
History lists > Saved Histories



Click on the history of interest (down arrowhead) and “Delete Permanently”.



Optionally, you can also transfer your data for example to a cloud service with Data transfer > Globus Data Transfer



Part I: Uploading data and starting the workflow

For this exercise (Part I), 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. Set the workflow to run overnight and view / interpret the results the next day (Part II).

Setting up your VectorBase Galaxy account

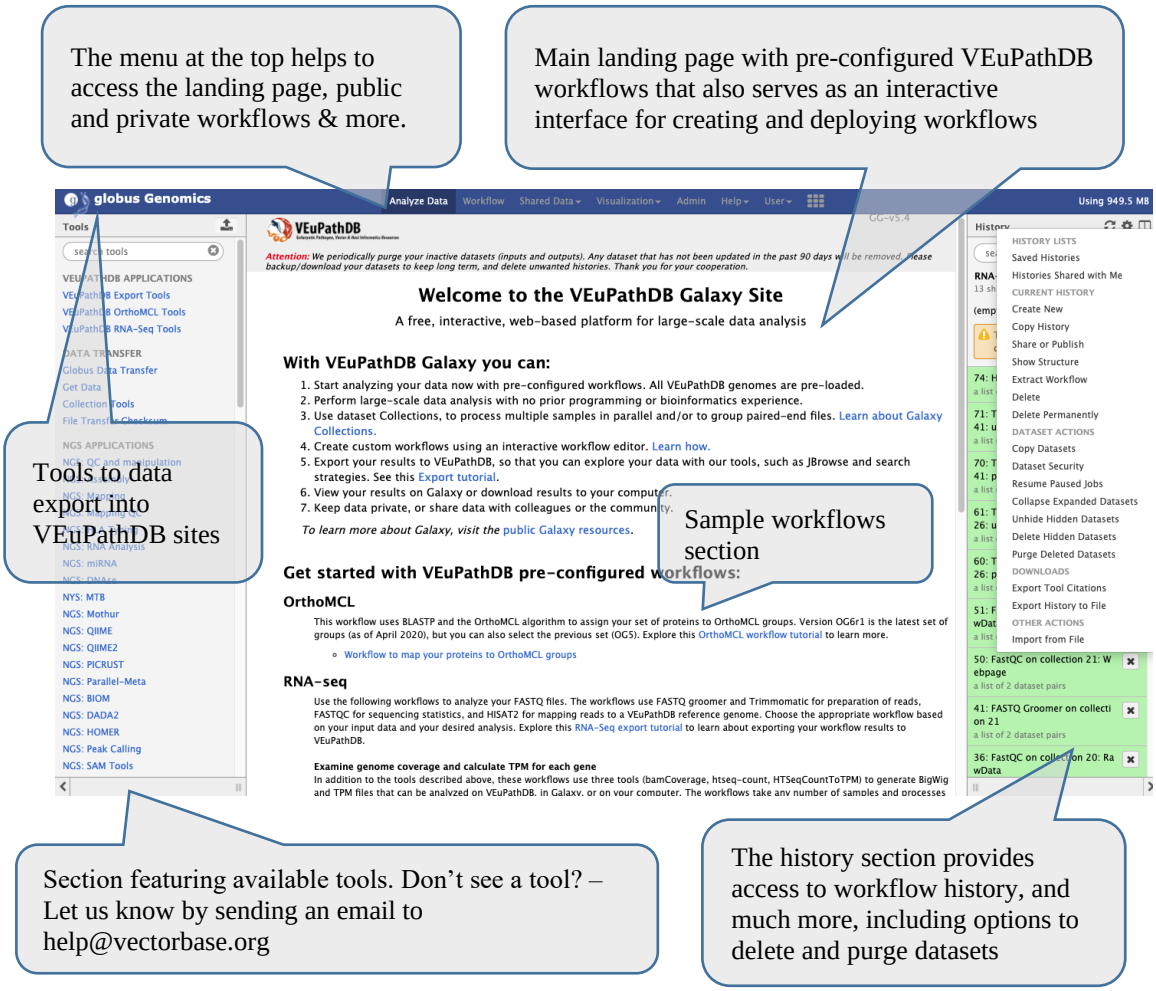
The step-by-step process has been explained in the other companion handout of this webinar.

The anatomy of the VectorBase 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. 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).

Sample data set:

We will be examining data from a study called "Transcriptome analysis of genes associated with pyrethroid resistance in South and Central American *Anopheles albimanus*". https://www.ncbi.nlm.nih.gov/bioproject/PRJNA498101.

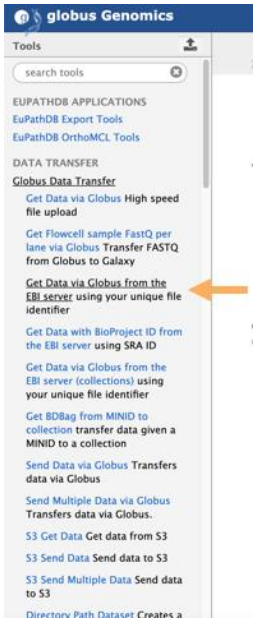
Sample: Accession Numbers (Name)	Sanarate	Guatemala	Peru
unexposed to insecticides	SAMN10341948 (San3) SAMN10341947 (San2) SAMN10341946 (San1)	SAMN10341945 (GTM_unx3) SAMN10341944 (GTM_unx2) SAMN10341943 (GTM_unx1)	SAMN10341939 (PER_unx3) SAMN10341938 (PER_unx2) SAMN10341937 (PER_unx1)
delta-methrin	-	SAMN10341942 (GTM_delta3) SAMN10341941 (GTM_delta2) SAMN10341940 (GTM_delta1)	SAMN10341933 (PER_delta3) SAMN10341932 (PER_delta2) SAMN10341931 (PER_delta1)
alpha-cypermethrin	-	-	SAMN10341936 (PER_acyp3) SAMN10341935 (PER_acyp2) SAMN10341934 (PER_acyp1)

Reference Genome in VectorBase: *Anopheles albimanus* STECLA AalbS2.7.

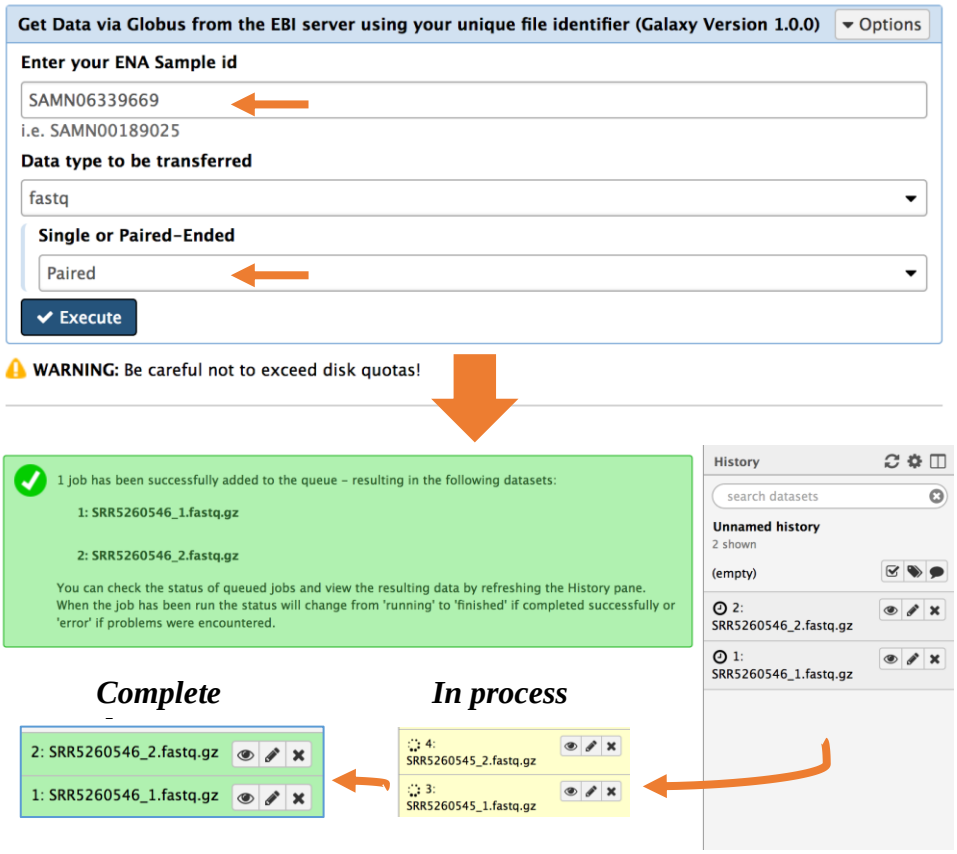
Comparisons: We will compare the San to the GTM-delta samples, you can later do the other comparisons following the same steps.

Condition-1	Condition-2
San	GTM-delta
San	GTM-unx
GTM-unx	GTM-delta
San	PER-delta
San	PER-unx
San	PER-acyp
PER-unx	PER-delta
PER-unx	PER-acyp
PER-delta	PER-acyp
GTM-delta	PER-delta

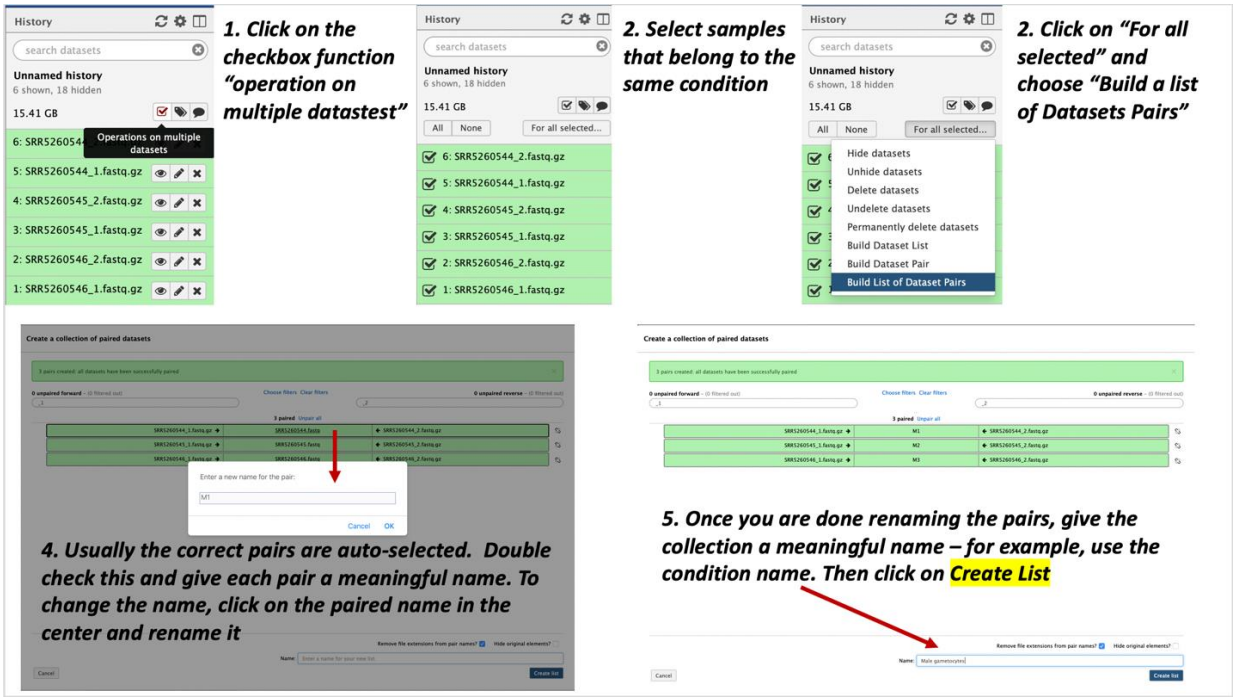
Step 1: Click on the "**Globus Data Transfer**" link in the left-hand menu. This will reveal a list of options; click on "**Get Data via Globus from the EBI server**". ***important: do not select the option for transferring a collection.



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.*



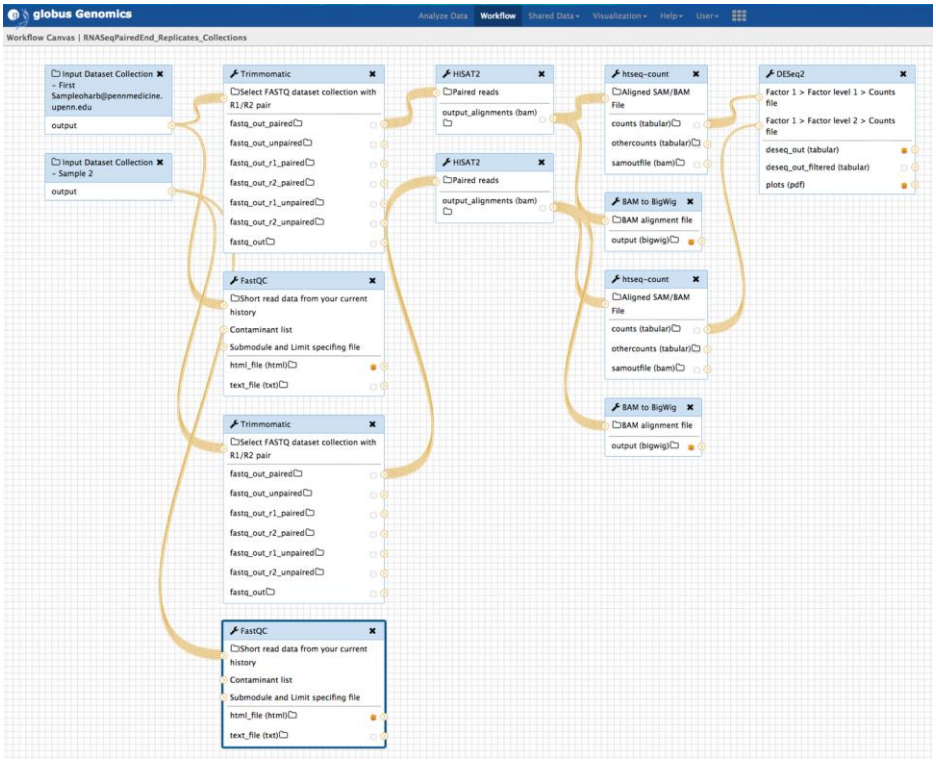
Step 3: If you are working with a dataset with biological replicates it is useful to organize the different conditions of your experiment into "Collections". For example, if your experiment included RNAseq from *Anopheles albimanus* susceptible laboratory colony Sanarate, San (three biological replicates, San1, San2, San3) and field collected mosquitoes from Guatemala and Peru (three biological replicates each, unexposed and exposed to two different insecticides), it is useful to organize these into collections. Using collections also reduces the complexity of the Galaxy workflows. 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:

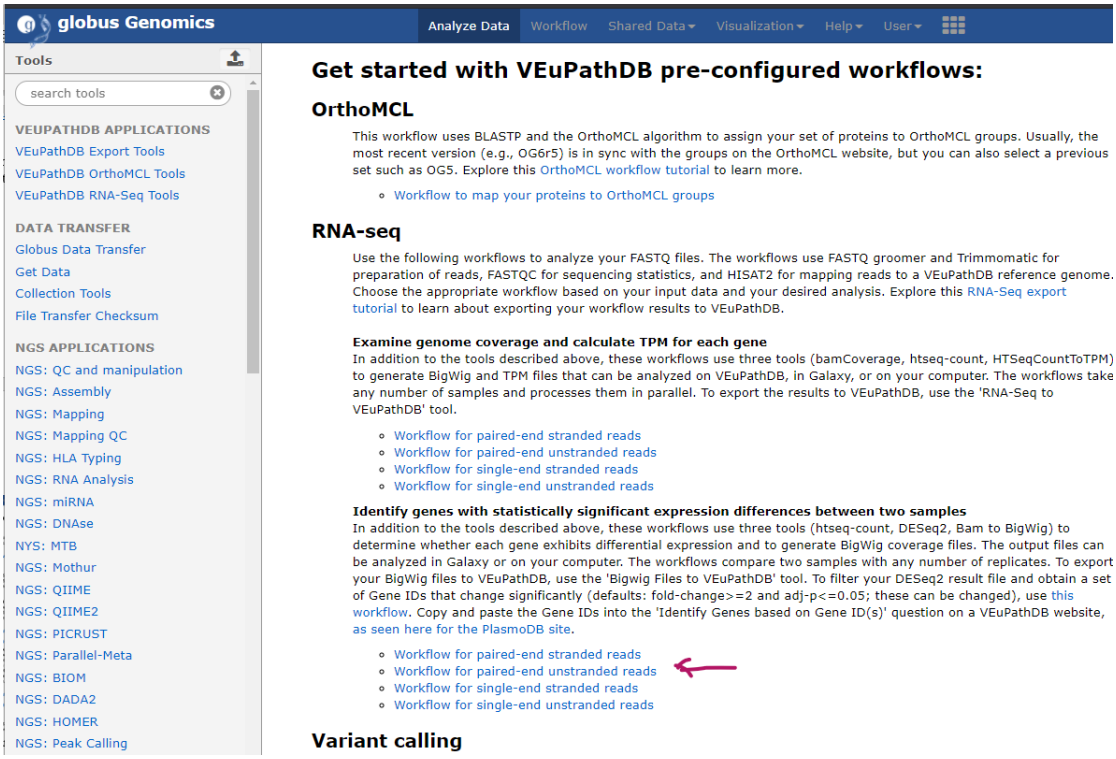
1. Analyzes the reads in your files and generates FASTQC reports.
2. Trims the reads based on their quality scores and adaptor sequences (Trimmomatic).
3. Aligns the reads to a reference genome using HISAT2 and generates coverage plots.
4. Determines read counts per gene (HTSeq)
5. Determines differential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



Additional resources:

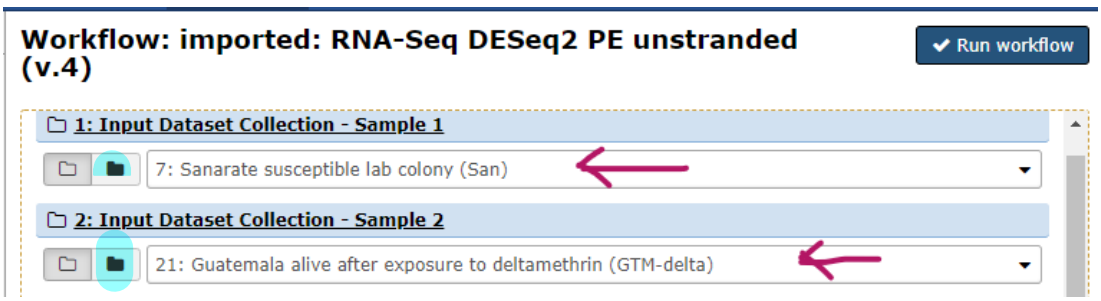
Galaxy Project: https://usegalaxy.org/
Trimmomatic: http://www.usadellab.org/cms/index.php?page=trimmomatic
FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
HISAT2: http://daehwankimlab.github.io/hisat2/main/
HTseq: https://htseq.readthedocs.io/en/master/
DEseq2: https://doi.org/10.1186/s13059-014-0550-8

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



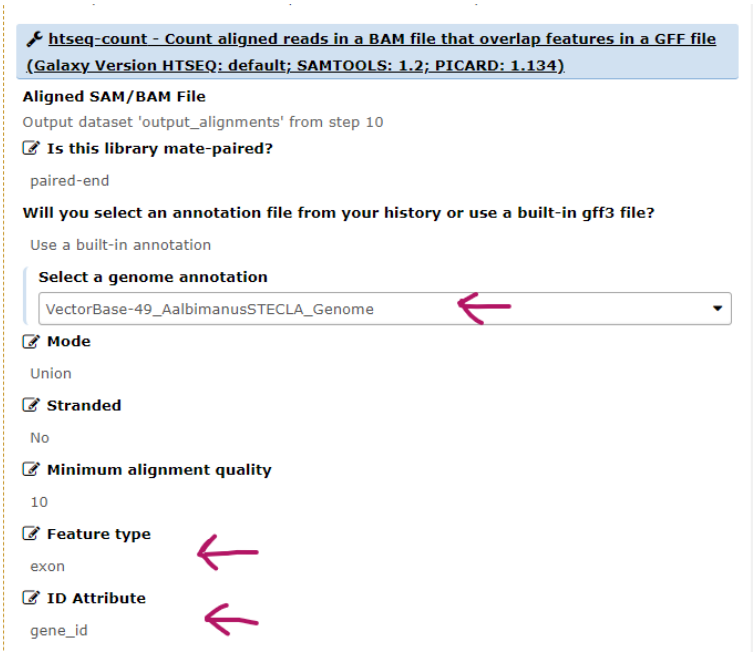
- Configure your workflow – there are multiple steps in the workflow, but you do not need to configure all of them. For the purpose of this exercise, you will need to configure the following:

a. 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.

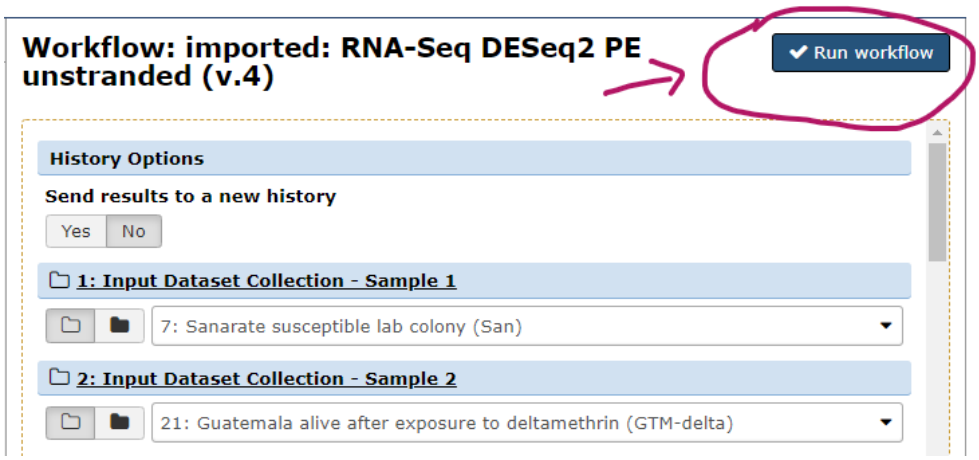


b. Some tools in the workflow require that you select the reference genome to be used. In this workflow, both HISAT2 and HTSeq require this (note that each of these tools is in the workflow twice since you have two collections). It is critical that you select the correct genome that matches the experimental organism. For this exercise, select the genome as shown in the image below

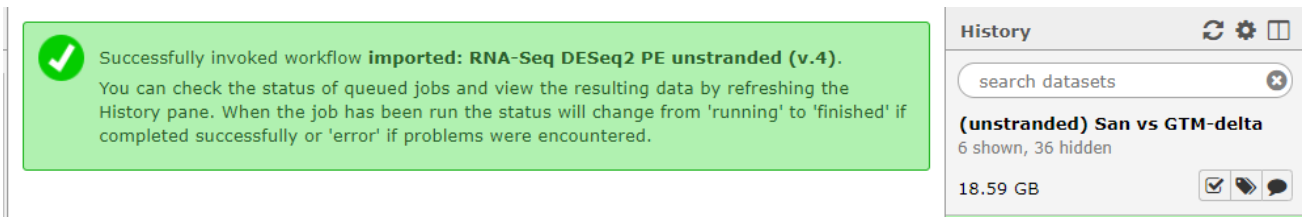
c. Two additional parameters to check in the htseq-count step are “Feature type” and “ID Attribute”. They should be set to “exon” and “gene_id”, respectively. Be aware that these are case-sensitive, so “Exon” is not correct but “exon” is correct. Here is how that step should look:



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



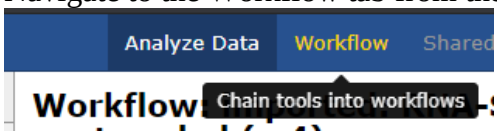
The steps will start running in the history section on the right. Grey means they are waiting to start. Yellow means they are running. Green means they have completed. Red means there was an error in the step.



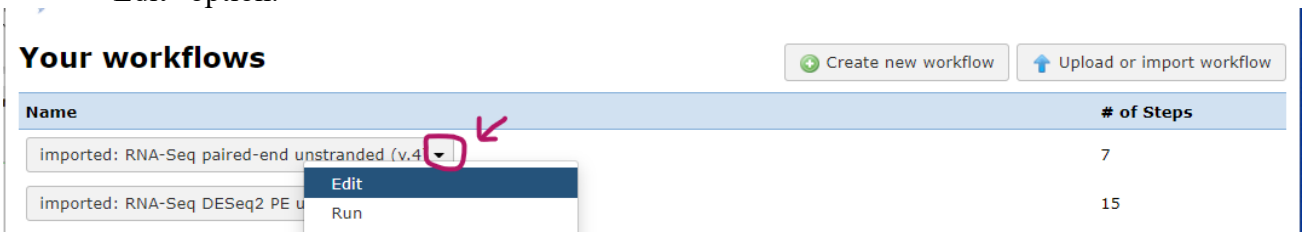
Practice working with Galaxy editor

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

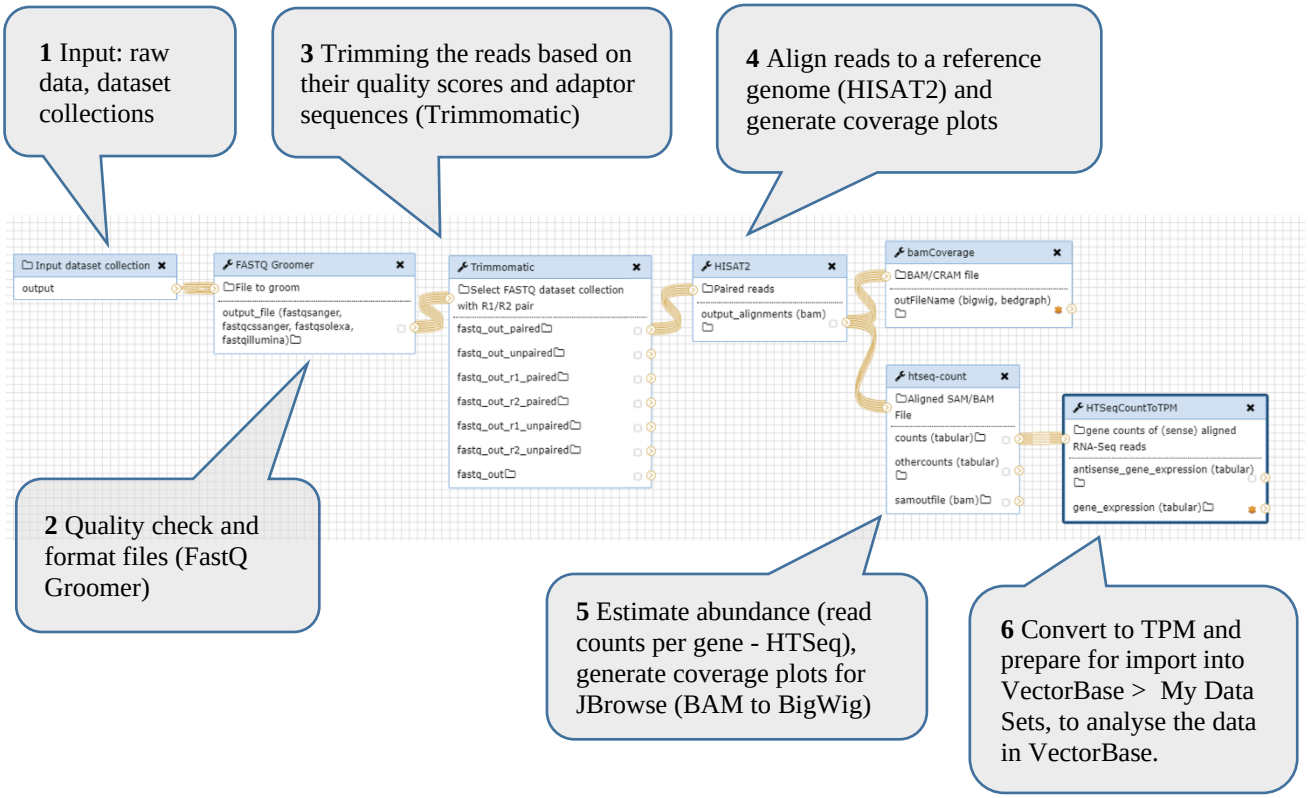
- Navigate to the Workflow tab from the main menu at the top



- Left click on the drop-down icon within the workflow you want to modify and select the "Edit" option.

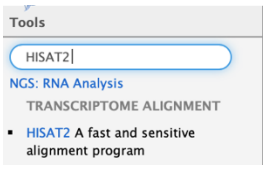


Sample workflow steps:



- Delete HISAT2 step by clicking on the " x " in the top right corner.

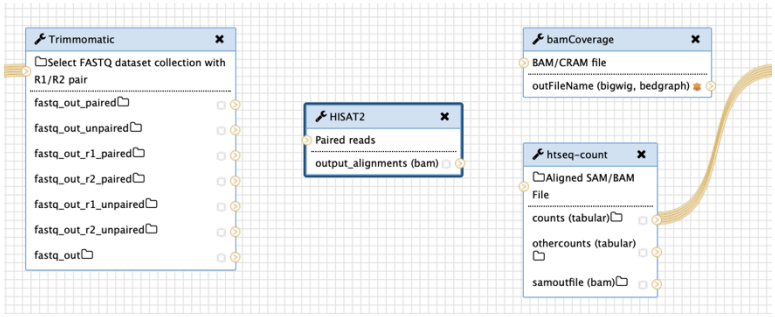
- Locate the HISAT2 tool in the Tools panel and click to insert it back into the workflow.



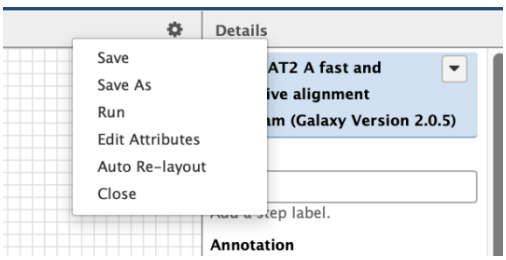
- Re-establish connections for HISAT2

- 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 if 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.

Appendix:

FASTQ files are text files (similar to FASTA) that include sequence quality information and details in addition to the sequence (ie. name, quality scores, sequencing machine ID, lane number etc.). FASTQ files are large and as a result not all sequencing repositories will store this format. However, tools are available to convert, for example, NCBI's SRA format to FASTQ. Sequence data is housed in three repositories that are synchronized on a regular basis.

- The sequence read archive (SRA) at GenBank
- The European Nucleotide Archive (ENA) at EMBL
- The DNA Data Bank of Japan