

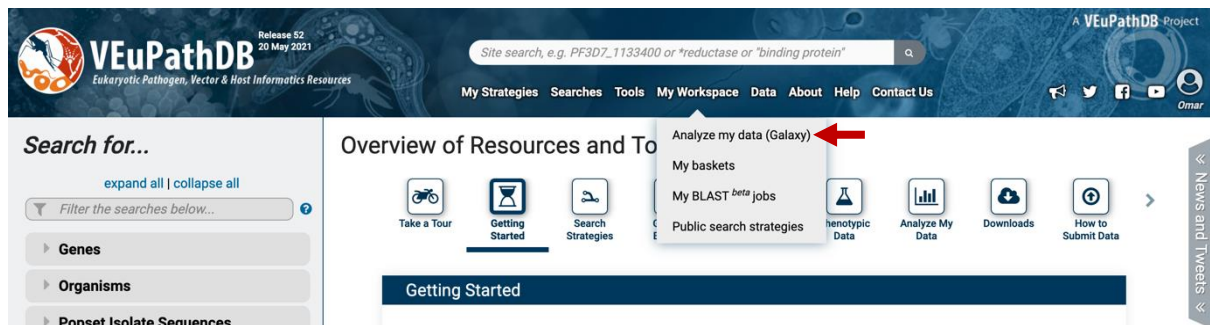
RNA sequence data analysis via VEuPathDB Galaxy, Part I Uploading data and starting the workflow (Group Exercise)

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.

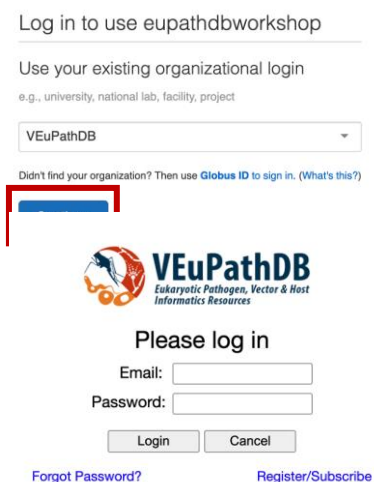
Section I: Setting up your VEuPathDB Galaxy account

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/>

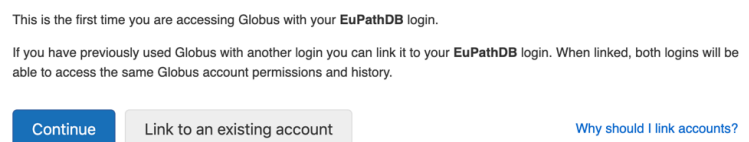
Step 2: On the next page you will be asked to define your organization. Choose VEuPathDB and click Continue.



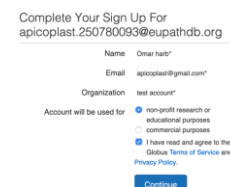
Step 3: If you are not already logged into VEuPathDB you will be prompted to do so now.

Welcome – You've Successfully Logged In

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”

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.

Step 5: Congratulations, you are in!

The anatomy of the VEuPathDB Galaxy landing page.

The workspace has four major components:

- the top menu controls the main interface
- the left panel has a list of available tools
- the main welcome page is the interactive interface that houses pre-configured workflows, workflows editor, etc.
- 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.

The screenshot shows the VEuPathDB Galaxy landing page. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The left sidebar lists various tools under categories like 'VEuPathDB APPLICATIONS', 'DATA TRANSFER', 'NCS APPLICATIONS', and 'NYS: MTB'. The main content area features a 'Welcome to the VEuPathDB Galaxy Site' message, a list of 'With VEuPathDB Galaxy you can:' actions, and sections for 'Get started with VEuPathDB pre-configured workflows' (OrthoMCL and RNA-seq). The right sidebar shows 'HISTORY LISTS' and 'Dataset Actions'. Callouts point to the top menu, the left tools panel, the main welcome section, and the right history panel.

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

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

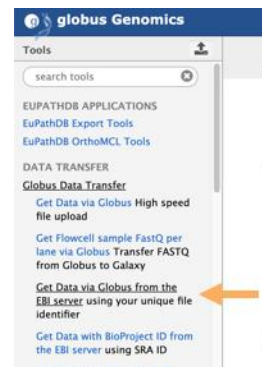
There are multiple ways to import data into your Galaxy workspace. For this exercise, we will use the 'Get Data via Globus from the EBI server using your unique file identifier' tool and enter the sequence repository sample IDs based on your group assignments (below). *Remember only one person in your 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.*

Get Data via Globus from the EBI server using your unique file identifier (Galaxy Version 1.0.0) Options

Enter your ENA Sample id
 ←
 i.e. SAMN00189025

Data type to be transferred

Single or Paired-Ended
 ←

⚠ **WARNING:** Be careful not to exceed disk quotas!

✓ 1 job has been successfully added to the queue – resulting in the following datasets:

1: SRR5260546_1.fastq.gz
 2: SRR5260546_2.fastq.gz

You can check the status of queued jobs and view the resulting data by refreshing the History pane. When the job has been run the status will change from 'running' to 'finished' if completed successfully or 'error' if problems were encountered.

Complete

2: SRR5260546_2.fastq.gz

1: SRR5260546_1.fastq.gz

In process

4: SRR5260545_2.fastq.gz

3: SRR5260545_1.fastq.gz

History

search datasets

Unnamed history
2 shown

(empty)

2: SRR5260546_2.fastq.gz

1: SRR5260546_1.fastq.gz

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:

Your uploaded data

Groups 1-3, PbANKA gametocyte data
6 shown
5.28 GB

6: SRR5260544_2.fastq.gz	👁️✎️✕
5: SRR5260544_1.fastq.gz	👁️✎️✕
4: SRR5260545_2.fastq.gz	👁️✎️✕
3: SRR5260545_1.fastq.gz	👁️✎️✕
2: SRR5260546_2.fastq.gz	👁️✎️✕
1: SRR5260546_1.fastq.gz	👁️✎️✕

Operations on multiple datasets

Build List of Dataset Pairs

Create a collection of paired datasets

3 pairs created: all datasets have been successfully paired

0 unpaired forward - (0 filtered out) Choose filters Clear filters

1

SRR5260544_1.fastq.gz →	erythrocyte3	SRR5260544_2.fastq.gz	✕
SRR5260545_1.fastq.gz →	erythrocyte2	SRR5260545_2.fastq.gz	✕
SRR5260546_1.fastq.gz →	erythrocyte1	SRR5260546_2.fastq.gz	✕

Name: erythrocytes

It's good to name the individual samples of the collection as well as the collection.

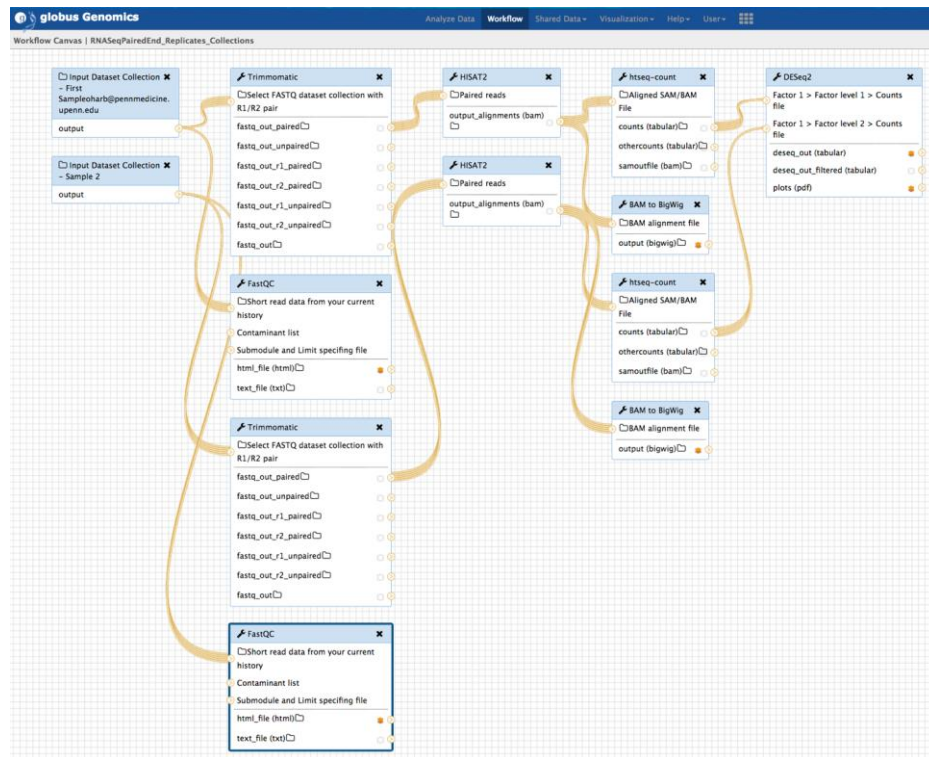
7: erythrocytes
a list of pairs with 3 items

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 \(https://usegalaxy.org/\)](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 is 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).

globus Genomics

Analyze Data | Workflow | Shared Data | Visualization | Help | User | Using 1.1 TB

Tools

search tools

VEUPATHDB APPLICATIONS

VEuPathDB Export Tools
VEuPathDB OrthoMCL Tools
VEuPathDB RNA-Seq Tools

DATA TRANSFER

Globus Data Transfer

Get Data via Globus High speed file upload
Get Flowcell sample FastQ per lane via Globus Transfer FASTQ from Globus to Galaxy
Get Data via Globus from the EBI server using your unique file identifier
Get Data with BioProject ID from the EBI server using SRA ID
Get Data via Globus from the EBI server (collections) using your unique file identifier
Get BDAG from MINID to collection transfer data given a MINID to a collection
Send Data via Globus Transfers data via Globus
Send Multiple Data via Globus Transfers data via Globus
S3 Get Data Get data from S3
S3 Send Data Send data to S3
S3 Send Multiple Data Send data to S3

Get started with VEuPathDB pre-configured workflows:

OrthoMCL

This workflow uses BLASTP and the OrthoMCL algorithm to assign your set of proteins to OrthoMCL groups. Version OG6r1 is the latest set of groups (as of April 2020), but you can also select the previous set (OG5). Explore this [OrthoMCL workflow tutorial](#) to learn more.

- Workflow to map your proteins to OrthoMCL groups

RNA-seq

Use the following workflows to analyze your FASTQ files. The workflows use FASTQ groomer and Trimmomatic for preparation of reads, FASTQC for sequencing statistics, and HISAT2 for mapping reads to a VEuPathDB reference genome. Choose the appropriate workflow based on your input data and your desired analysis. Explore this [RNA-Seq export tutorial](#) to learn about exporting your workflow results to VEuPathDB.

Examine genome coverage and calculate TPM for each gene

In addition to the tools described above, these workflows use three tools (bamCoverage, htseq-count, HTSeqCountToTPM) to generate BigWig and TPM files that can be analyzed on VEuPathDB, in Galaxy, or on your computer. The workflows take any number of samples and processes them in parallel. To export the results to VEuPathDB, use the 'RNA-Seq to VEuPathDB' tool.

- Workflow for paired-end stranded reads
- Workflow for paired-end unstranded reads
- Workflow for single-end stranded reads
- Workflow for single-end unstranded reads

Identify genes with statistically significant expression differences between two samples

In addition to the tools described above, these workflows use three tools (htseq-count, DESeq2, Bam to BigWig) to determine whether each gene exhibits differential expression and to generate BigWig coverage files. The output files can be analyzed in Galaxy or on your computer. The workflows compare two samples with any number of replicates. To export your BigWig files to VEuPathDB, use the 'BigWig Files to VEuPathDB' tool. To filter your DESeq2 result file and obtain a set of Gene IDs that change significantly (defaults: fold-change=2 and adj-p<0.05; these can be changed), use this [workflow](#). Copy and paste the Gene IDs into the 'Identify Genes based on Gene IDs' question on a VEuPathDB website, as seen here for the

- Workflow for paired-end stranded reads
- Workflow for paired-end unstranded reads
- Workflow for single-end stranded reads
- Workflow for single-end unstranded reads

Variant calling

Use the following workflows to analyze your FASTQ files. The workflows use Sickle for preparation of reads, Bowtie2 for mapping reads to a VEuPathDB reference genome, Freebayes for variant detection, SnpEff to evaluate the effect of variants, and SnpSift for filtering.

History

search datasets

tgm test
13 shown

7.18 GB

13: BAM to BigWig on collection 5
6
a list of 2 datasets

12: BAM to BigWig on data
55
a list of 2 datasets

11: BAM to BigWig on data
54
a list of 2 datasets

10: BAM to BigWig on collection 5
3
a list of 2 datasets

9: BAM to BigWig on data
52
a list of 2 datasets

8: BAM to BigWig on data
51
a list of 2 datasets

7: Test TPMs for Eve
a list of 4 datasets

6: HTSeqCountToTPM on collection 71: gene expression
a list of 2 datasets

5: FC16
a list of 2 datasets

4: FC6
a list of 2 datasets

3: HTSeqCountToTPM on collection 65: gene expression
a list of 2 datasets

Workflow: imported: DESeq2 Workflow for paired-end unstranded reads (v.7) (imported from uploaded file)

Run Workflow

History Options

Send results to a new history

Yes No

1: Input Dataset Collection - Sample 1

13: dsGFP_infected

2: Input Dataset Collection - Sample 2

14: dsISARI_infected

Source for the reference genome to align against

Use a built-in genome

Select a reference genome

VectorBase-49_IscaipularisWiki_Genome

If your genome of interest is not listed, contact the Galaxy team

Factor level

1: Factor level

Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'

Counts file(s)

2: Factor level

Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'

Counts file(s)

Factor level

1: Factor level

Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'

dsGFP_infected

Only letters, numbers and underscores in this field

Counts file(s)

2: Factor level

Specify a factor level, typical values could be 'tumor', 'normal', 'treated' or 'control'

Sample2

Counts file(s)

- 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* iANKA, the reference genome you select should be *PlasmoDB-51_PbergheiANKA_Genome* (**C above figure**).

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

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.

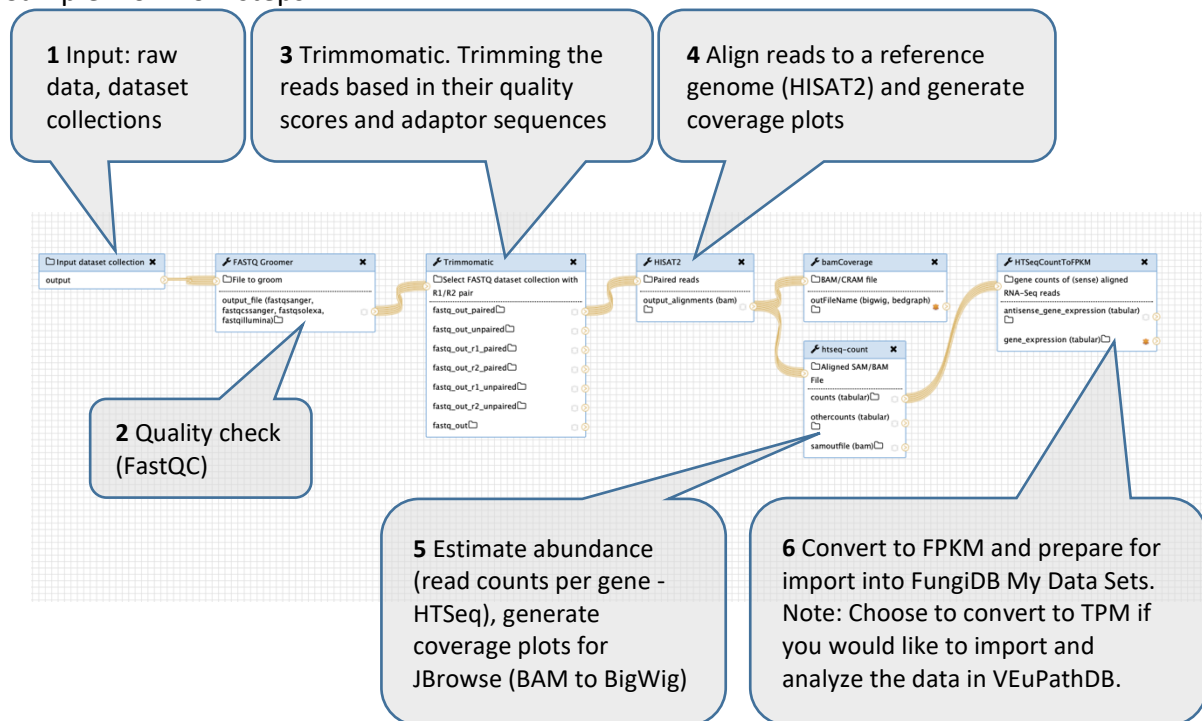
8

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.

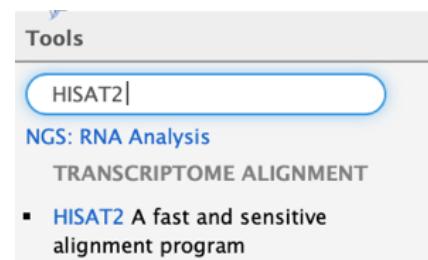
- Navigate to the Workflow tab from the main menu at the top and select
- Left click on the drop-down icon within the workflow you want to modify and select the “Edit” option.

Sample workflow steps:

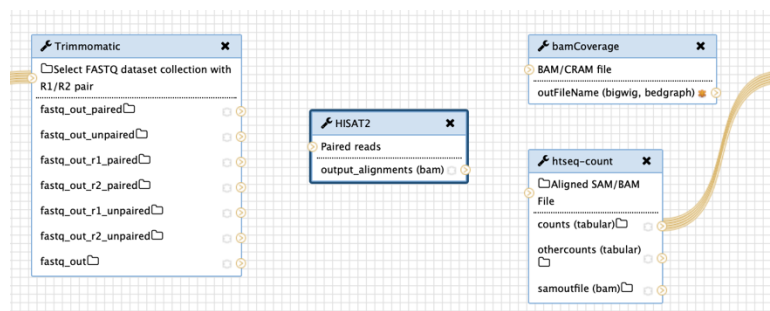


Delete HISAT2 step by clicking on the “ x ” in the top right corner in the workflow.

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

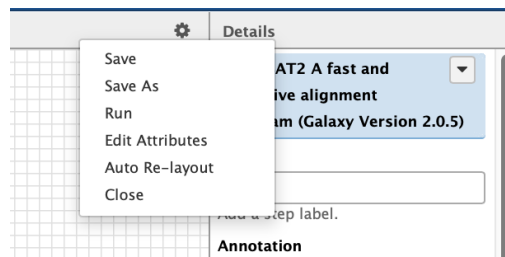


- Re-establish connections for HISAT2. Click on the arrow in the step before HISAT2 and drag to the appropriate input in HISAT2 tool.



- 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.