

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

The goal of this exercise is to use a Galaxy workflow to analyze RNA sequencing data. 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. EuPathDB 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 4-6 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 EuPathDB Galaxy account

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

<http://eupathdbworkshop.globusgenomics.org/>

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

globus Globus Account Log In

Log in to use EupathDBWorkshop

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

EuPathDB

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

Continue

Or

Sign in with Google Sign in with ORCID ID

Step 3: Log in to EuPathDB (if you are not logged in already).



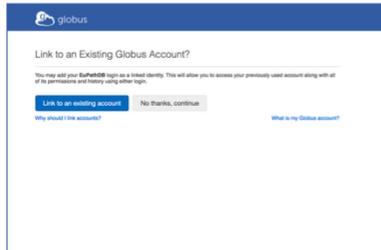
Step 4: Next, sign up for the EuPathDB Galaxy instance.

Analyze My Experiment

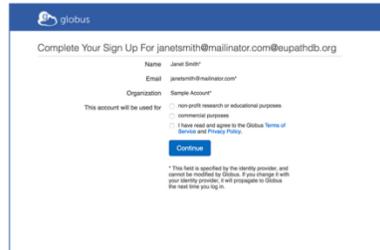
The first time you visit EuPathDB Galaxy you will be asked to sign up with Globus, EuPathDB's Galaxy instance manager. This is a three-step sign-up process (screenshots below).

Click **"Continue to Galaxy"** to sign up for EuPathDB Galaxy services.

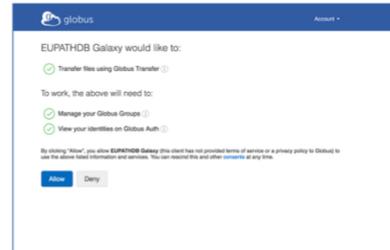
Contact us if you experience any difficulties.



(1) If you already have a Globus account, you can link it to your EuPathDB account. **Your choice.** If you don't have a prior Globus account, choose **No Thanks**.



(2) Complete your account information and agree to Globus's Terms and Conditions. Please read, make your selections, and click **Continue**.

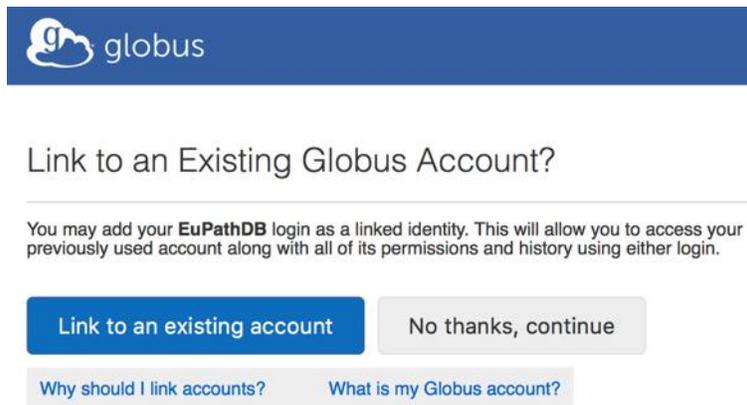


(3) Grant permission to share your Globus identity and files with us. Please click **Allow**. (We will only perform file transfers that you explicitly request, between Galaxy and other resources, including EuPathDB.)

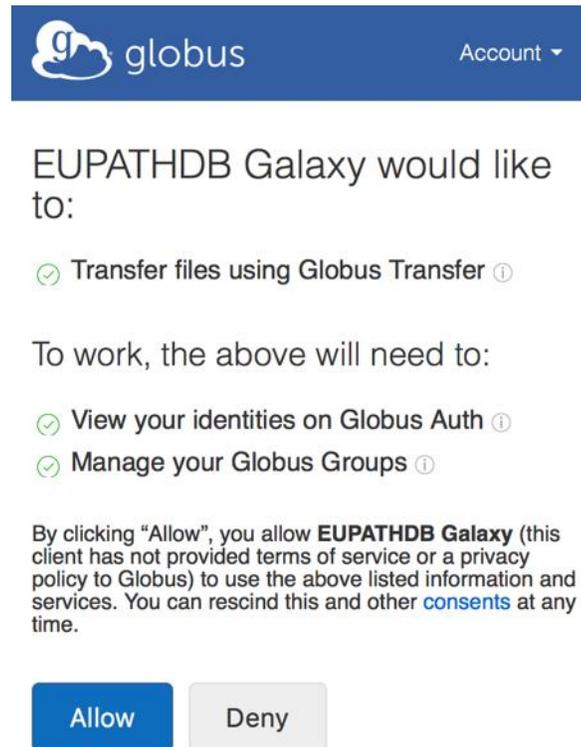
[Continue to Galaxy](#)

Step 5: Click on "Continue to Galaxy" and follow the instructions.

Step 6: Click on "No thanks, continue"



Step 7: Click on “Allow”



Step 8: Congratulations, you are in!

Section II: Importing data to Galaxy

There are multiple ways to important data into your Galaxy workspace. For this exercise, we will use the ‘Download from web or upload from disk’ tool and enter the direct data repository links listed below under ‘Group Assignments’. Remember one person in your group will be starting 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 don’t want to overload the servers. The samples below were all generated by paired end sequencing, hence there are two files for each sample. The files are fastq files that are compressed (that is why they end in .gz = gzip).

Step 1: Click on the “Get data” link in the left-hand menu. This will reveal a list of options; click on “**Get Data via Globus from the EBI server**”

The screenshot shows the EuPathDB Galaxy interface. On the left, the 'Tools' menu is open, and 'Get Data via Globus from the EBI server using your unique file identifier' is selected, indicated by an orange arrow. The main area displays a 'Welcome to the EuPathDB Galaxy Site' message and a list of instructions for using the platform. Below this, there are sections for 'Get started with pre-configured workflows:' including 'OrthoMCL' and 'RNA Sequencing'. The right-hand panel shows a 'History' section with a list of datasets, including SRR6855530_2.fastq.gz, SRR6855530_1.fastq.gz, SRR6855527_2.fastq.gz, SRR6855527_1.fastq.gz, SRR685852_2.fastq.gz, SRR685852_1.fastq.gz, SRR685423_2.fastq.gz, SRR685423_1.fastq.gz, SRR6854541_2.fastq.gz, SRR6854541_1.fastq.gz, SRR6854543_2.fastq.gz, and SRR6854543_1.fastq.gz.

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

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

Enter your ENA Sample id

SAMN06339669 

i.e. SAMN00189025

Data type to be transferred

fastq

Single or Paired-Ended

Paired 

 **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.

History refresh settings close

search datasets clear

Unnamed history
2 shown

(empty) check refresh close

2: SRR5260546_2.fastq.gz view edit close

1: SRR5260546_1.fastq.gz view edit close

Completed

2: SRR5260546_2.fastq.gz view edit close

1: SRR5260546_1.fastq.gz view edit close

In process

4: SRR5260545_2.fastq.gz view edit close

3: SRR5260545_1.fastq.gz view edit close

Queued for import



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 *Plasmodium falciparum* male gametocyte stages (three biological replicates) and erythrocytic stages (three biological replicates), it is useful to organize these into two collections, one that includes all male gametocyte files and the other that includes all the erythrocytic stage files. Using collections also reduces the complexity of the Galaxy workflows. See below:

1. Click on the checkbox function "operation on multiple datasetst"

2. Select samples that belong to the same condition

2. Click on "For all selected" and choose "Build a list of Datasets Pairs"

4. Usually the correct pairs are auto-selected. Double check this and give each pair a meaningful name. To change the name, click on the paired name in the center and rename it

5. Once you are done renaming the pairs, give the collection a meaningful name – for example, use the condition name. Then click on **Create List**

Group assignments:

We will divide up into groups and use a Galaxy workflow to analyze RNAseq data from a study that examined gene expression in different *Toxoplasma* developmental stages: “An experimental genetically attenuated live vaccine to prevent transmission of *Toxoplasma gondii* by cats” <https://www.ncbi.nlm.nih.gov/pubmed/30728393>

The data is available in the sequence repositories:

<https://www.ebi.ac.uk/ena/data/view/PRJNA428456>

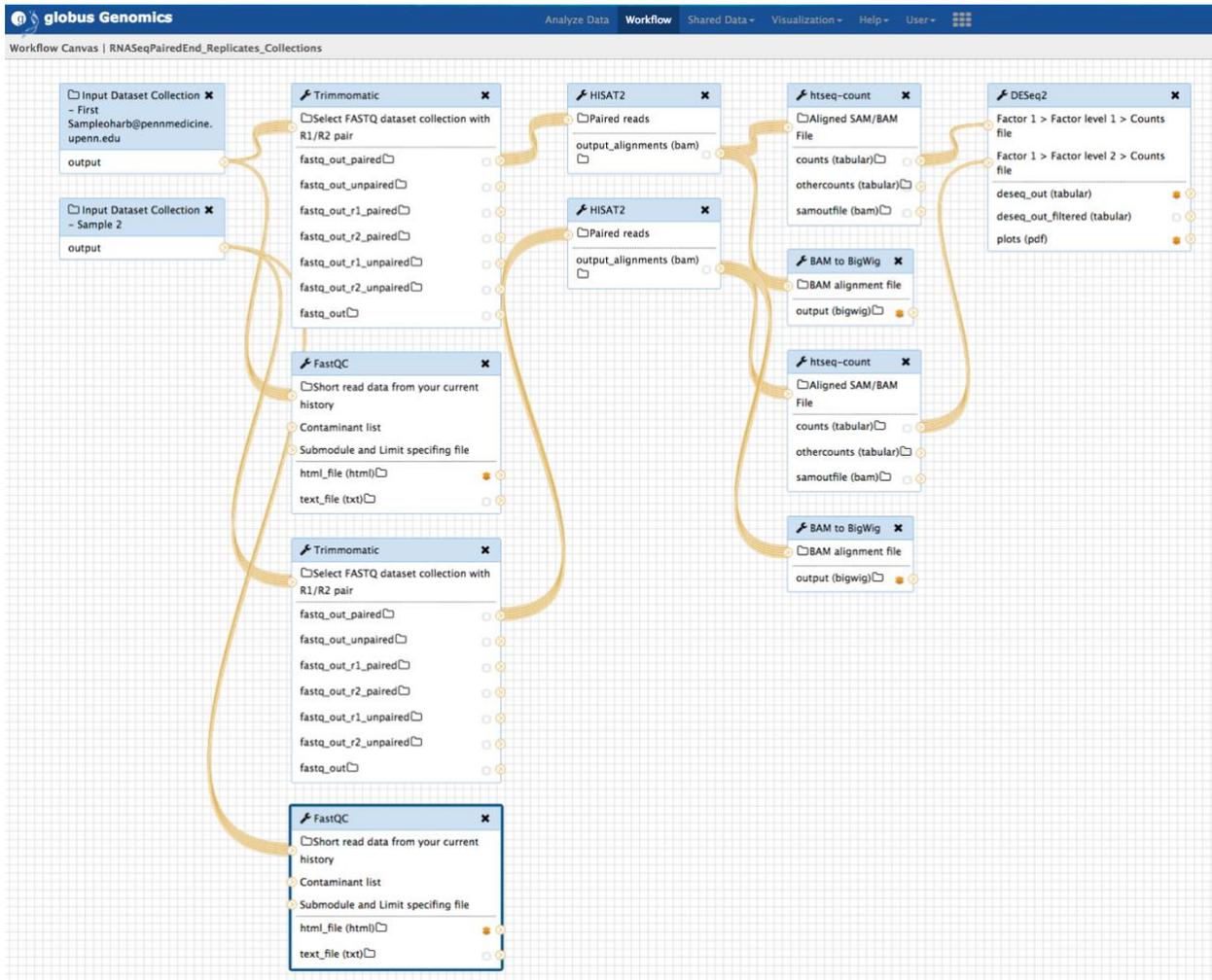
Sample Name	Cat Day 3	Cat Day 5	Cat Day 7	Cysts	Tachyzoites
Sample	SAMN08291340	SAMN08291334	SAMN08291331	SAMN08291324	SAMN08291326
Accession	SAMN08291338	SAMN08291333	SAMN08291329	SAMN08291323	SAMN08291325
Numbers	SAMN08291336	SAMN08291332	SAMN08291327		

Group Number	1	2	3	4	5	6
Comparison	Day 3 vs. Day 5	Day 3 vs. Day 7	Day 5 vs. Day 7	Cysts vs. Day 7	Tachyzoites vs. Day 7	Tachyzoites vs. Cysts

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).



- To use one of the EuPathDB preconfigured workflows, go to the Galaxy home page and select the workflow that you would like to run. For this exercise “**EuPathDB Workflow for Illumina paired-end RNA-seq, biological replicates**” – click on this workflow to run it

The screenshot shows the Galaxy Genomics interface. The main content area is titled "With EuPathDB Galaxy you can:" and lists several tasks:

- Start analyzing your data now. All EuPathDB genomes are pre-loaded. Pre-configured workflows are available.
- Perform large-scale data analysis with no prior programming or bioinformatics experience.
- Create custom workflows using an interactive workflow editor. [Learn how](#)
- Visualize your results (BigWig) in GBrowse.
- Keep data private, or share it with colleagues or the community.

 Below this, there is a link: "To learn more about Galaxy check out public Galaxy resources: [Learn Galaxy](#)".

 The "Get started with pre-configured workflows:" section lists:

- OrthoMCL**: Map your proteins to OrthoMCL groups. This workflow uses BLASTP and the OrthoMCL algorithm to assign your set of proteins to OrthoMCL groups. [Explore this tutorial to learn more.](#)
- RNA Sequencing**:
 - EuPathDB RNA-Seq paired-end: for RNAseq Export Tool**: This workflow generates BigWig and Expression files that are compatible with the EuPathDB RNAseq Export Tool. [Explore this tutorial to learn more.](#)
 - EuPathDB Workflow for Illumina paired-end RNA-seq, without replicates**: Profile a transcriptome and analyze differential gene expression. Tools: FastQC, Sickle, GSNAP, Cufflinks, CuffDiff.
 - EuPathDB Workflow for Illumina paired-end RNA-seq, without replicates**: Profile a transcriptome and analyze differential gene expression. Tools: FastQC, Trimmomatic, TopHat2, Cufflinks, CuffDiff.
 - EuPathDB Workflow for Illumina paired-end RNA-seq, biological replicates**: Profile a transcriptome and analyze differential gene expression. Tools: FastQC, Trimmomatic, TopHat2, HTseq, DESeq2. (This item is highlighted with a red arrow.)
 - EuPathDB Workflow for Illumina paired-end RNA-seq, biological replicates**: Profile a transcriptome and analyze differential gene expression. Tools: Collections, FastQC, Trimmomatic, HISAT2, HTseq, DESeq2.
- Variant Calling**: [EuPathDB Workflow for Variant Calling: detect and annotate](#)

 On the right side, the "History" panel shows a list of datasets, including "24: Female to Male", "23: Female to RBCs", "22: Male to RBCs", "21: Female Gametocytes", "20: SRR5260550_2.fastq.gz", "19: SRR5260550_1.fastq.gz", "18: SRR5260551_2.fastq.gz", "17: SRR5260551_1.fastq.gz", "16: SRR5260552_2.fastq.gz", "15: SRR5260552_1.fastq.gz", "14: Male Gametocytes", "13: SRR5260547_2.fastq.gz", and "12: SRR5260547_1.fastq.gz".

- 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.

The screenshot shows the configuration page for the workflow "RNASeqPairedEnd_Replicates_Collections". The "Run workflow" button is visible in the top right. The "History Options" section has "Send results to a new history" set to "No". The "Input dataset collection" section shows two collections selected:

- 1: Input dataset collection - 1: 13: Erythrocytic Stages
- 2: Input dataset collection - 13: 18: Male Gametocytes

 The workflow steps are:

- 3: Trimmomatic - 3 (Galaxy Version 0.36.5)
- 4: FastQC - 2 (Galaxy Version FASTQC: 0.11.3)
- 5: Trimmomatic - 9 (Galaxy Version 0.36.5)
- 6: FastQC - 8 (Galaxy Version FASTQC: 0.11.3)
- 7: HISAT2 - 4 (Galaxy Version 2.0.5)

 The "Input data format" section is set to "FASTQ". Under "Single end or paired reads?", "Paired reads" is selected. The "Source for the reference genome to align against" is set to "Use a built-in genome". The "Select a reference genome" dropdown is set to "AmoebaDB-29_AstronyxisUnknown_Genome".

- 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 these tools are 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*, the reference genome you select should be *Plasmodium berghei*.

Source for the reference genome to align against

Use a built-in genome

Select a reference genome

- PlasmoDB-29_Pchabaudichabaudi_Genome
- PlasmoDB-29_PcynomolgiB_Genome
- PlasmoDB-29_Pfalciparum3D7_Genome
- PlasmoDB-29_PfalciparumIT_Genome
- PlasmoDB-29_PknowlesiH_Genome
- PlasmoDB-29_PreichenowiCDC_Genome
- PlasmoDB-29_PvivaxP01_Genome
- PlasmoDB-29_PvivaxSal1_Genome
- PlasmoDB-29_Pyoeliiyoelii17XNL_Genome
- PlasmoDB-29_PyoeliiyoeliiYM_Genome
- PlasmoDB-30_PcoatneyiHackeri_Genome
- PlasmoDB-30_PfragileNilgiri_Genome
- PlasmoDB-30_PinuiSanAntonio1_Genome
- PlasmoDB-30_PmalariaeUG01_Genome
- PlasmoDB-30_PvinckeipetteriCR_Genome
- PlasmoDB-30_Pvinckeivinckeivincke_Genome
- PlasmoDB-30_Pyoeliiyoelii17X_Genome
- PlasmoDB-32_PbergheiANKA_Genome
- PlasmoDB-32_PgallinaceumBA_Genome
- PlasmoDB-32_PovalecurtisiGH01_Genome

Paired alignment parameters

Use default values

- c. Another very important parameter to check in the htseq-count step is the Feature type. The default is usually set to exon. Make sure you change this to gene. To change this to gene, click on the edit icon, the type the word “gene”. This is case sensitive so be careful about this.

htseq-count – Count aligned reads in a BAM file that overlap features in a GFF file (Galaxy Version HTSEQ: default; SAMTOOLS: 1.2; PICARD: 1.134)

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

Select a genome annotation

PlasmoDB-32_PbergheiANKA_Genome

Mode

Union

Stranded

Yes

Minimum alignment quality

0

Feature type

gene

Feature type (3rd column in GFF file) to be used. All features of other types are ignored. The default, suitable for RNA-Seq and Ensembl GTF files, is exon.

ID Attribute

ID

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

The screenshot displays the Globus Genomics web interface. At the top, the navigation bar includes 'g globus Genomics', 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main content area is divided into three sections:

- Tools (Left):** A sidebar with a search bar and a list of applications categorized under 'EUPATHDB APPLICATIONS' and 'NGS APPLICATIONS'. The NGS applications include QC and manipulation, Assembly, Mapping, Mapping QC, RNA Analysis, DNase, Mothur, QIIME, PICRUST, Parallel-Meta, BIOM, HOMER, Peak Calling, SAM Tools, and various versions of GATK Tools.
- Message (Center):** A green notification box with a checkmark icon. The text reads: 'Successfully invoked workflow RNASeqPairedEnd_Replicates_Collections. 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.'
- History (Right):** A panel titled 'History' showing a list of datasets for the project 'Cultured vs. Salivary'. The list includes:
 - 28: FastQC on data 4: RawData (Grey)
 - 27: FastQC on data 4: Webpage (Grey)
 - 26: FastQC on data 3: RawData (Grey)
 - 25: FastQC on data 3: Webpage (Grey)
 - 24: FastQC on data 2: RawData (Grey)
 - 23: FastQC on data 2: Webpage (Grey)
 - 22: FastQC on data 1: RawData (Grey)
 - 21: FastQC on data 1: Webpage (Grey)
 - 20: Trimmomatic on collection 5: unpaired (Green)
 - 19: Trimmomatic on collection 5: paired (Green)
 - 10: Cultured sporozoites (Yellow)
 - 5: Sporozoites (Yellow)

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.

