

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

Additional resources:

[Galaxy Project \(https://usegalaxy.org/\)](https://usegalaxy.org/)

[Trimmomatic manual](#)

[FastQC](#)

[HISAT2](#)

[HTseq](#)

[DEseq2](#)

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.

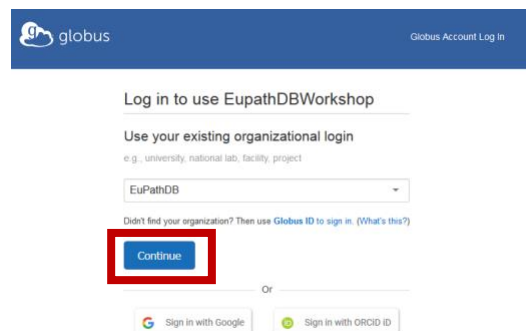
We will be working in groups. Each group will have 4-6 members. One person in the group will run the Galaxy controls on one computer. The other members' roles are to ensure that the correct datasets are used and that the correct workflow parameters are selected.

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.



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



Please log in

Email:

Password:

[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.

[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

[REDACTED]@eupathdb.org

Name **[REDACTED]**

Email **[REDACTED]**

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](#).

* 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.

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 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:

Groups 1, 2 & 3 will be examining data from a study called “*Plasmodium berghei* transcriptome for female gametocytes, male gametocytes, and asexual erythrocytic stages”

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5604118/>

The data is available in the sequence repositories:

<https://www.ebi.ac.uk/ena/data/view/PRJNA374918>

Sample Name	Erythrocyte stages (Asexual)	Male gametocytes	Female gametocytes
Sample Accession Numbers	SAMN06339669 SAMN06339670 SAMN06339671	SAMN06339666 SAMN06339667 SAMN06339668	SAMN06339663 SAMN06339664 SAMN06339665

Group Number	1	2	3
Comparison	Erythrocyte stages vs. Male gametocytes	Erythrocyte stages vs. Female gametocytes	Male gametocytes vs. Female gametocytes

Groups 4, 5 & 6 will be examining data from a study called “*Plasmodium falciparum* NF54 Transcriptome” which examines RNAseq from 3 stages: erythrocytic, salivary gland and cultured sporozoite stages. This study is unpublished but data is accessible in the sequence repositories:

<https://www.ebi.ac.uk/ena/data/view/PRJNA230379>

Sample Name	Erythrocyte stages (Asexual)	Salivary gland sporozoites	Cultured sporozoites
Sample Accession Numbers	SAMN02428730 SAMN02428734	SAMN02428726 SAMN02428729	SAMN02428728 SAMN02428727

Group Number	4	5	6
Comparison	Erythrocyte stages vs. Salivary gland sporozoites	Erythrocyte stages vs. Cultured sporozoites	Salivary gland sporozoites vs. Cultured sporozoites

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.*

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.

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:

- 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 EuPathDB Galaxy interface. On the left is a sidebar with navigation links. The main panel is titled 'With EuPathDB Galaxy you can:' and lists several workflow categories: OrthoMCL, RNA Sequencing, and Variant Calling. Under RNA Sequencing, the workflow 'EuPathDB Workflow for Illumina paired-end RNA-seq, biological replicates' is highlighted with a red arrow. The right sidebar shows a 'History' panel with a list of dataset pairs.

- 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.

Workflow: RNASeqPairedEnd_Replicates_Collections Run workflow

History Options

Send results to a new history
☐ Yes ☒ No

1: Input dataset collection - 1
☐ 13: Erythrocytic Stages

2: Input dataset collection - 13
☐ 18: Male Gametocytes

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)

Input data format
 FASTQ

Single end or paired reads?
 Collection of paired reads

Paired reads
 Output dataset 'fastq_out_paired' from step 3

Paired-end options
 Use default values

Source for the reference genome to align against
 Use a built-in genome

Select a reference genome
 AmoebaDB-29_AastronyxisUnknown_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_Pyoelliyoelii17XNL_Genome
- ☐ PlasmoDB-29_PyoelliyoeliiYM_Genome
- ☐ PlasmoDB-30_PcoatneyiHackeri_Genome
- ☐ PlasmoDB-30_PfragileNilgiri_Genome
- ☐ PlasmoDB-30_PinuiSanAntonio1_Genome
- ☐ PlasmoDB-30_PmalariaeUG01_Genome
- ☐ PlasmoDB-30_PvinckeipetteriCR_Genome
- ☐ PlasmoDB-30_Pvinckeivinckeivincke1_Genome
- ☐ PlasmoDB-30_Pyoelliyoelii17X_Genome
- ☒ PlasmoDB-32_PbergheiANKA_Genome
- ☐ PlasmoDB-32_Pgallinaceum8A_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, then 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.

globo**us** Genomics

Analyze Data
Workflow
Shared Data
Visualization
Help
User

Using 381.1 GB

Tools

[Get Data](#)

EUPATHDB APPLICATIONS

EuPathDB Export Tools

NGS APPLICATIONS

NGS: QC and manipulation
 NGS: Assembly
 NGS: Mapping
 NGS: Mapping QC
 NGS: RNA Analysis
 NGS: DNase
 NGS: Mothur
 NGS: QIIME
 NGS: PICRUST
 NGS: Parallel-Meta
 NGS: BIOM
 NGS: HOMER
 NGS: Peak Calling
 NGS: SAM Tools
 NGS: SAM Tools (1.1)
 NGS: BAM Tools
 NGS: SNPIR Tools
 NGS: Picard
 NGS: Picard (1.128)
 NGS: Picard (2.7.1)
 NGS: Indel Analysis
 NGS: GATK Tools
 NGS: GATK2 Tools
 NGS: GATK3 Tools
 NGS: GATK3 Tools (3.6)
 NGS: GATK3 Tools (3.8)

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

Cultured vs. Salivary

53 shown, 8 hidden

9.7 GB

28: FastQC on data 4: RawData

27: FastQC on data 4: Webpage

26: FastQC on data 3: RawData

25: FastQC on data 3: Webpage

24: FastQC on data 2: RawData

23: FastQC on data 2: Webpage

22: FastQC on data 1: RawData

21: FastQC on data 1: Webpage

20: Trimmomatic on collection 5: unpaired

19: Trimmomatic on collection 5: paired

10: Cultured sporozoites

5: Sporozoites

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.

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.

- FASTA**
- Definition line
- >SEQUENCE_1
- MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDK
AVQLLREKGLGKAAKKADRLAAEGLVSVKVSDDFTIAA
MRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRL
KDPNKPEHKIPQFASRKQLSDAILKEAEEKIKEELKAQ
GKPEKIWDNIIPGKMNSFIADNSQLDSKLTLMGQFYVM
DDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKT
EDFAAEVAAQL
- Sequence
- FASTQ**
- End of Sequence
- @SRR016080.2 20AKUAAXX:7:1:123:268
TGTAGCATAATGCCGTTTCTTTGTTCCATTCATC
+
||&I&4|C|||||||.|||3:||||3#6||||1|)
@SRR016080.3 20AKUAAXX:7:1:112:638
TATAGATCTTGGAACACCCGTTGTATTATTCGCAA
+
|||||
@SRR016080.4 20AKUAAXX:7:1:102:360
TTGCCAGTACAACACCGTTTTGCATCGTTTTTTT
+
|||||\$|||||'|||||@||||D35
- Definition line
- Sequence
- Encoded Quality Score