

RNA sequence data analysis in VEuPathDB Galaxy, Part I

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. The EuPathDB project, now known as VEuPathDB, developed its Galaxy instance in collaboration with Globus Genomics. To learn how to use Galaxy, follow this link to access tutorials prepared by the Galaxy Training Network: https://wiki.galaxyproject.org/Learn#Galaxy_101

Learning objectives:

1. [Retrieve raw sequence data from the sequence repository EBI using Globus Data Transfer tool;](#)
2. [Run an RNA-Seq workflow for paired-end reads.](#)

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 genome, calculate expression values and determine differential expression.

You will need to have a VEuPathDB account to use VEuPathDB Galaxy services. If you don't have an account, click on "Register" at the top right corner of the page to set up a free account. The username and password will work on any VEuPathDB site.

The screenshot shows the EuPathDB Project website. The header includes the EuPathDB logo, the text "Release 46 6 Nov 2019", and the "EuPathDB Project" label. A search bar contains "Gene ID: PF3D7_1133400" and "Gene Text Search: synth". Navigation links include "Home", "New Search", "My Strategies", "My Basket (0)", "Tools", "Data Summary", "Downloads", "Community", "Analyze My Experiment", and "My Favorites". A banner for a webinar reads: "Register for our upcoming webinar: Running a Galaxy workflow and integrating data into VEuPathDB. Thursday May 14th at 10AM US Eastern time." Below the banner, the "Data Summary" section features a "News and Tweets" sidebar with updates from November 2019. The main content area states: "The EuPathDB Bioinformatics Resource Center provides a portal for accessing genomic-scale datasets associated with the diverse eukaryotic microbes (mouse-over the following logos for information on component websites):". It then displays logos for various microbial databases: AmoebaDB, CryptoDB, FungiDB, GiardiaDB, MicrosporidiaDB, PiroplasmaDB, PlasmoDB, ToxoDB, TrichDB, TriTrypDB, and OrthoMCL.

Once you have an account, follow the steps outlined in the "Setting up your EuPathDB Galaxy account" tutorial to get started.

1. Retrieve raw sequence data from the sequence repository EBI using Globus Data Transfer tool;

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

We will examine data from a study called “*Plasmodium berghei* transcriptome for female gametocytes, male gametocytes, and asexual erythrocytic stages”

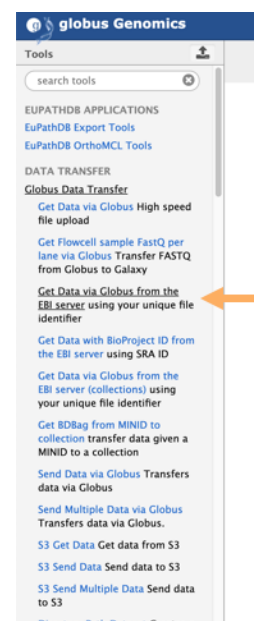
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5604118/>

The data is available in the sequence repositories:

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

Sample Name	Erythrocyte stages (Asexual)	Male gametocytes	Comparison
Sample Accession Numbers	SAMN06339669 SAMN06339670 SAMN06339671	SAMN06339666 SAMN06339667 SAMN06339668	Erythrocyte stages vs. Male gametocytes

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.

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 ←

Execute

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

search datasets

Unnamed history

2 shown

(empty)

2: SRR5260546_2.fastq.gz

1: SRR5260546_1.fastq.gz

Comple

2: SRR5260546_2.fastq.gz

1: SRR5260546_1.fastq.gz

In

4: SRR5260545_2.fastq.gz

3: SRR5260545_1.fastq.gz

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**

2. Running a workflow in VEuPathDB 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 VEuPathDB preconfigured workflows, go to the VEuPathDB Galaxy home page and select the workflow that you would like to run. For this exercise, we will use **"Workflow for paired-end unstranded reads"** – click on this workflow to run it

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 coverage across the genome and calculate RPKM for each gene

In addition to the tools described above, these workflows use three tools (bamCoverage, htseq-count, HTSeqCountToFPKM) to generate BigWig and FPKM files that can be analyzed on VEuPathDB, in Galaxy, or on your computer. The workflows take any number of samples and process the samples in parallel.

- 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. NOTE: Export of DESeq2 results to VEuPathDB will be available soon.

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

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:

Workflow: RNASeqPairedEnd_Replicates_Collections

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_AstronysisUnknown_Genome

- 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.
- 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. 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_PvinckeivinckeiVinckei_Genome

PlasmoDB-30_Pyoeliiyoelii17X_Genome

PlasmoDB-32_PberghelANKA_Genome

PlasmoDB-32_Pgallinaceum8A_Genome

PlasmoDB-32_PvalecurtisiGH01_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.

Workflow: RNASeqPairedEnd_Replicates_Collections

✓ Run workflow

RNASeqPairedEnd_Replicates_Collections (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

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.

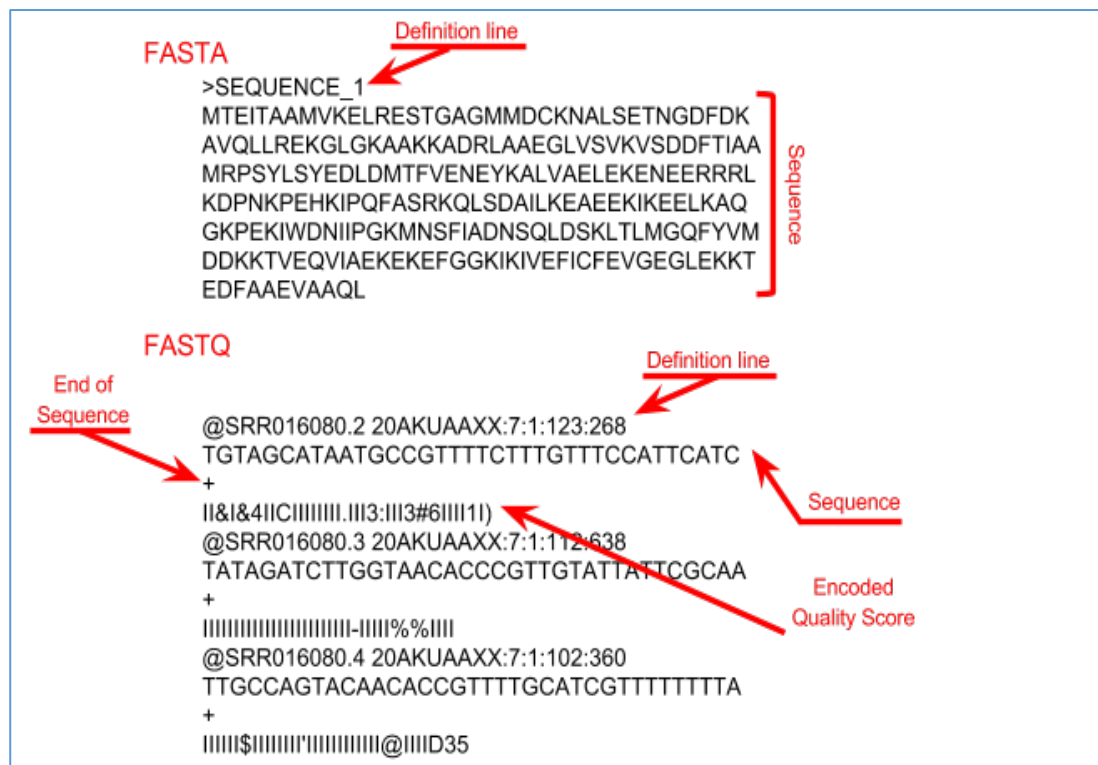
The screenshot displays the 'globus Genomics' interface. The top navigation bar includes 'Analyze Data', 'Workflow', 'Shared Data', 'Visualization', 'Help', and 'User'. The main area is divided into three panels:

- Tools (Left):** A sidebar with a search bar and a list of applications categorized under 'EUPATHDB APPLICATIONS' and 'NGS APPLICATIONS'. The 'NGS APPLICATIONS' list includes various tools like QC, Assembly, Mapping, RNA Analysis, DNase, Mothur, QIIME, PICRUST, Parallel-Meta, BIOM, HOMER, Peak Calling, SAM Tools, SNP Tools, Picard, Indel Analysis, GATK Tools, and GATK3 Tools.
- Message (Center):** A green notification box with a checkmark icon stating: '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 jobs. The jobs are color-coded: grey for 'waiting', yellow for 'running', green for 'completed', and red for 'error'. The jobs listed are:
 - 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 (red)
 - 19: Trimmomatic on collection 5: paired (red)
 - 10: Cultured sporozoites (red)
 - 5: Sporozoites (red)

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 at GenBank
- The European Nucleotide Archive at EMBL
- The DNA data bank of Japan



Additional resources (tool manuals):

[Trimmomatic](#)

[FastQC](#)

[HISAT2](#)

[HTseq](#)

[DEseq2](#)