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/)
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.

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”

Step 5: Congratulations, you are in!
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

<table>
<thead>
<tr>
<th>Sample Name</th>
<th>Erythrocyte stages (Asexual)</th>
<th>Male gametocytes</th>
<th>Female gametocytes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Accession Numbers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMN06339669</td>
<td>SAMN06339666</td>
<td>SAMN06339663</td>
<td></td>
</tr>
<tr>
<td>SAMN06339670</td>
<td>SAMN06339667</td>
<td>SAMN06339664</td>
<td></td>
</tr>
<tr>
<td>SAMN06339671</td>
<td>SAMN06339668</td>
<td>SAMN06339665</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group Number</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>Erythrocyte stages vs. Male gametocytes</td>
<td>Erythrocyte stages vs. Female gametocytes</td>
<td>Male gametocytes vs. Female gametocytes</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Sample Name</th>
<th>Erythrocyte stages (Asexual)</th>
<th>Salivary gland sporozoites</th>
<th>Cultured sporozoites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Accession Numbers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAMN02428730</td>
<td>SAMN02428726</td>
<td>SAMN02428728</td>
<td></td>
</tr>
<tr>
<td>SAMN02428734</td>
<td>SAMN02428729</td>
<td>SAMN02428727</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group Number</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comparison</td>
<td>Erythrocyte stages vs. Salivary gland sporozoites</td>
<td>Erythrocyte stages vs. Cultured sporozoites</td>
<td>Salivary gland sporozoites vs. Cultured sporozoites</td>
</tr>
</tbody>
</table>
**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 *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:
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.

• 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 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*.

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.

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.
Appendix:

FASTQ file 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