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.

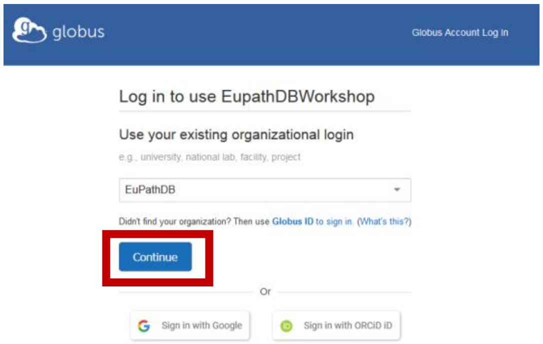
We will be working in groups. Each group will have 4-6 members. One person in the group will run the Galaxy controls on one computer. The other members' roles are to ensure that the correct datasets are used and that the correct workflow parameters are selected.

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.



Log in to use EupathDBWorkshop

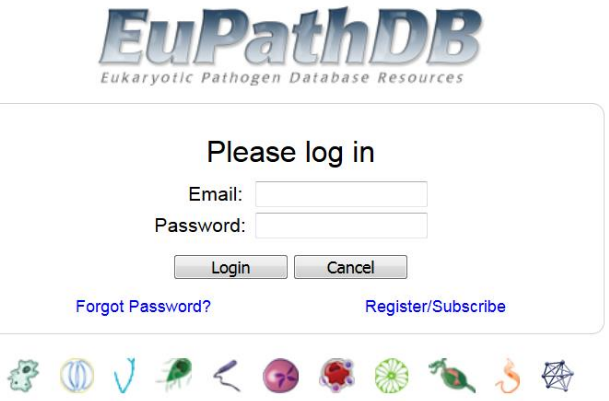
Use your existing organizational login

e.g., university, national lab, facility, project

EuPathDB

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

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



EuPathDB

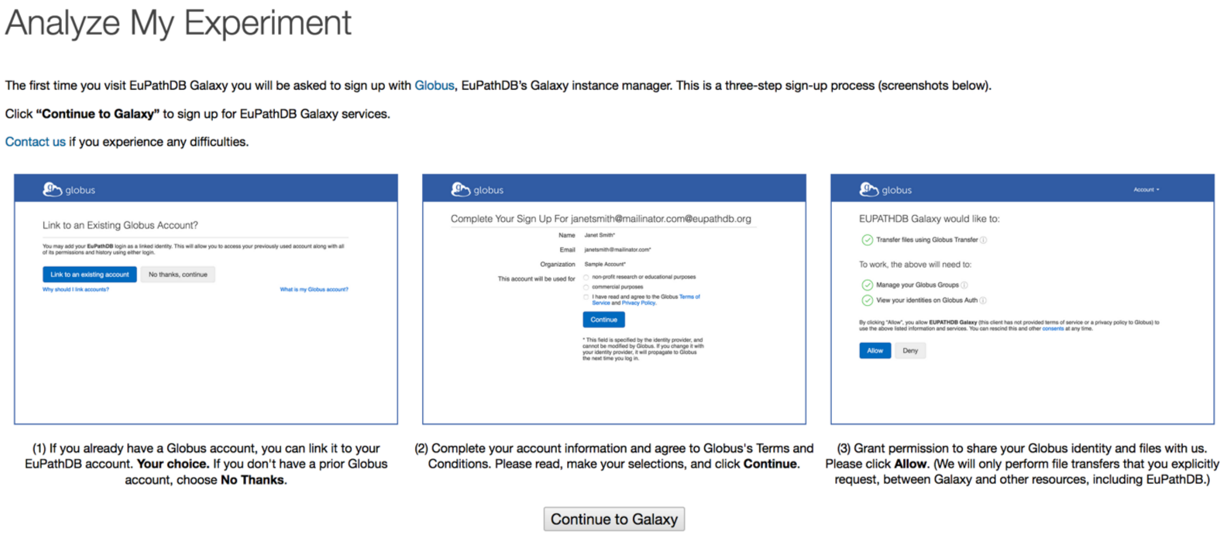
Eukaryotic Pathogen Database Resources

Please log in

Email:

Password:

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.

Link to an Existing Globus Account?

You may add your **EuPathDB** login as a linked identity. This will allow you to access your previously used account along with all of its permissions and history using either login.

globus
Link to an existing account
No thanks, continue
Why should I link accounts?
What is my Globus account?

Complete Your Sign Up For janetsmith@mailinator.com@eupathdb.org

Name Janet Smith*

Email janetsmith@mailinator.com*

Organization Sample Account*

This 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.

EUPATHDB Galaxy would like to:

✓ Transfer files using Globus Transfer

To work, the above will need to:

✓ Manage your Globus Groups

✓ View your identities on Globus Auth

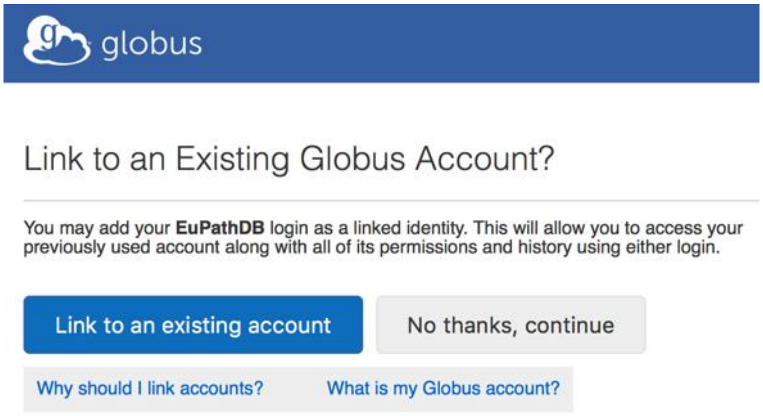
By clicking "Allow", you allow **EUPATHDB Galaxy** (this client has not provided terms of service or a privacy policy to Globus) to use the above listed information and services. You can rescind this and other consents at any time.

(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.)

Step 5: Click on "Continue to Galaxy" and follow the instructions.
Step 6: Click on "No thanks, continue"

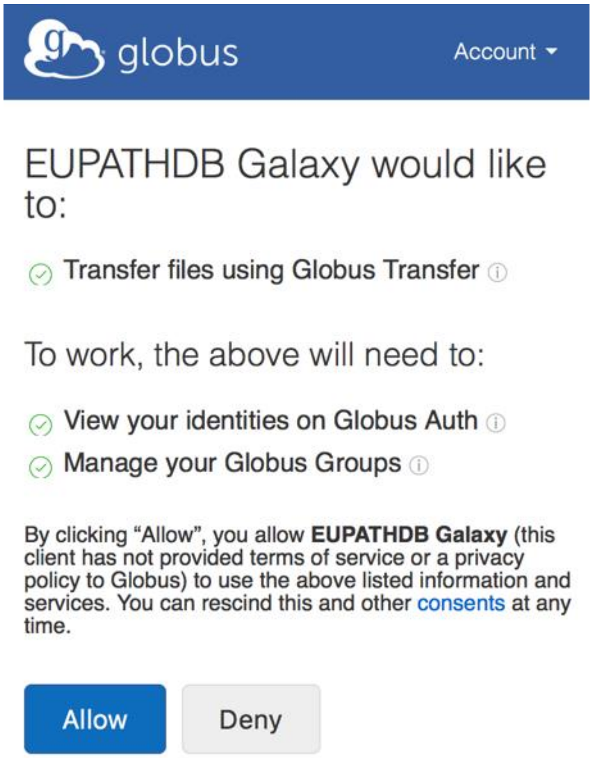


Link to an Existing Globus Account?

You may add your **EuPathDB** login as a linked identity. This will allow you to access your previously used account along with all of its permissions and history using either login.

globus
Link to an existing account
No thanks, continue
Why should I link accounts?
What is my Globus account?

Step 7: Click on “Allow”



EUPATHDB Galaxy would like to:

- Transfer files using Globus Transfer

To work, the above will need to:

- View your identities on Globus Auth
- Manage your Globus Groups

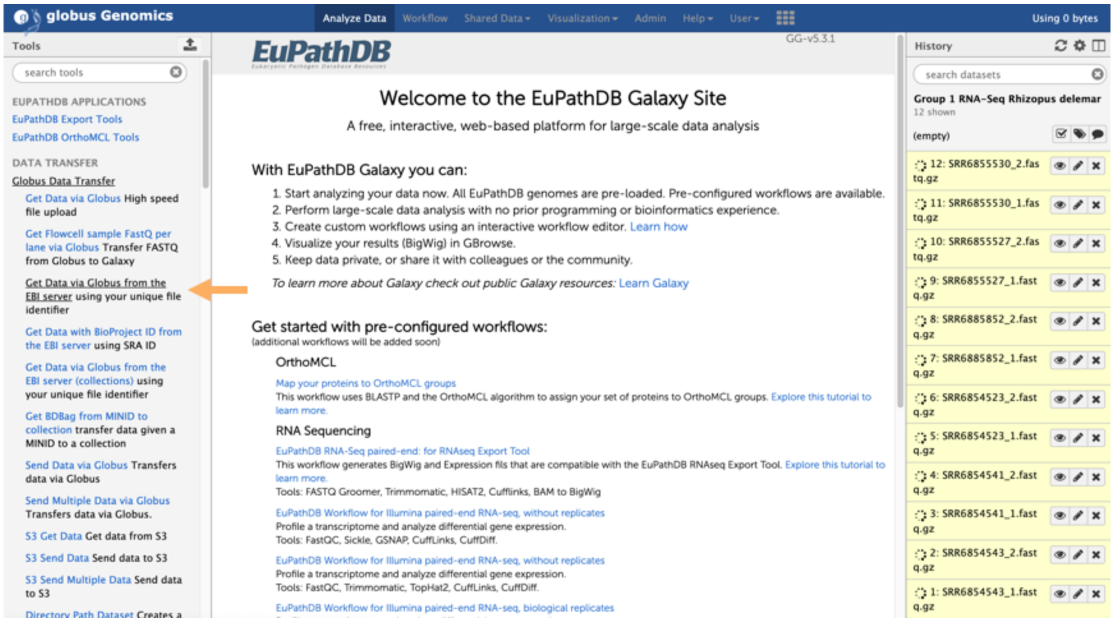
By clicking “Allow”, you allow **EUPATHDB Galaxy** (this client has not provided terms of service or a privacy policy to Globus) to use the above listed information and services. You can rescind this and other consents at any time.

Step 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**”



EUPATHDB APPLICATIONS

EuPathDB Export Tools

EuPathDB OrthoMCL 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 BDBag 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

Directory Path Dataset Creates a

Welcome to the EuPathDB Galaxy Site

A free, interactive, web-based platform for large-scale data analysis

With EuPathDB Galaxy you can:

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

To learn more about Galaxy check out public Galaxy resources: Learn Galaxy

Get started with pre-configured workflows:
(additional workflows will be added soon)

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 fils that are compatible with the EuPathDB RNAseq Export Tool. Explore this tutorial to learn more.
Tools: FASTQ Groomer, Trimmomatic, HISAT2, Cufflinks, BAM to BigWig

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.

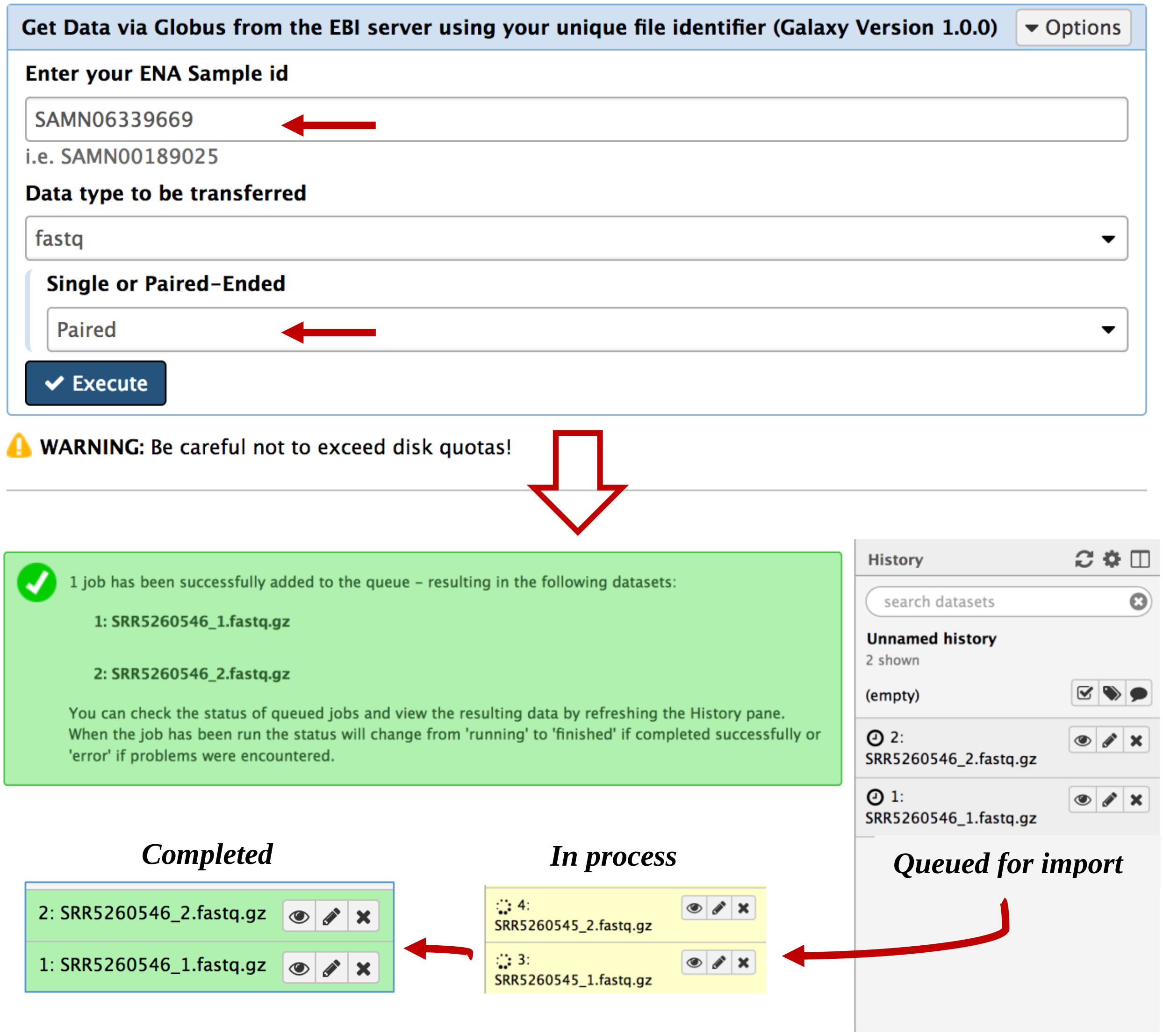
EuPathDB Workflow for Illumina paired-end RNA-seq, biological replicates

Group 1 RNA-Seq Rhizopus delemar
12 shown

(empty)

12: SRR6855530_2.fastq.gz
11: SRR6855530_1.fastq.gz
10: SRR6855527_2.fastq.gz
9: SRR6855527_1.fastq.gz
8: SRR6885852_2.fastq.gz
7: SRR6885852_1.fastq.gz
6: SRR6854523_2.fastq.gz
5: SRR6854523_1.fastq.gz
4: SRR6854541_2.fastq.gz
3: SRR6854541_1.fastq.gz
2: SRR6854543_2.fastq.gz
1: 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)

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

Unnamed history
2 shown
(empty)

2: SRR5260546_2.fastq.gz

1: SRR5260546_1.fastq.gz

Completed

2: SRR5260546_2.fastq.gz

1: SRR5260546_1.fastq.gz

In process

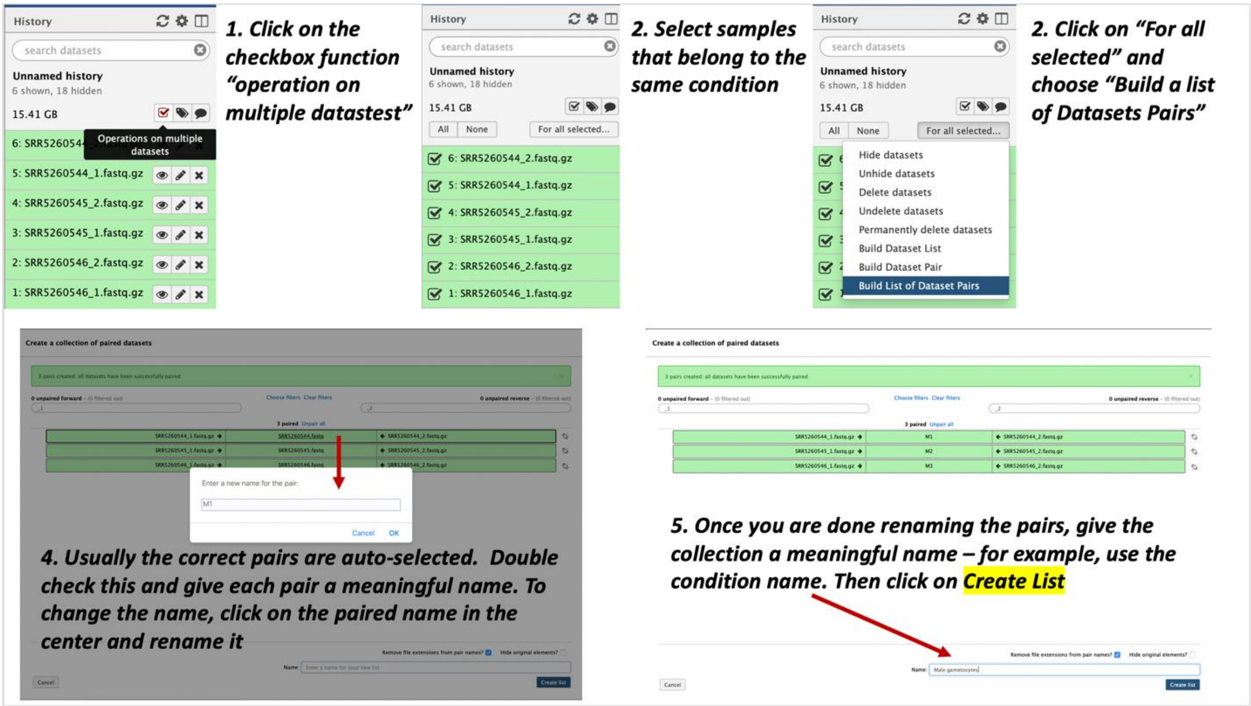
4: SRR5260545_2.fastq.gz

3: SRR5260545_1.fastq.gz

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 datastest"

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***

Enter a new name for the pair: M1

Name: Male gametocytes

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

Samples:

Erythrocyte stages (Asexual):
SAMN06339669
SAMN06339670
SAMN06339671

Male gametocytes:
SAMN06339666
SAMN06339667
SAMN06339668

Female gametocytes:
SAMN06339663
SAMN06339664
SAMN06339665

Group 1:

Plasmodium berghei male gametocytes vs. erythrocytic stages

Group 2:

Plasmodium berghei female gametocytes vs. erythrocytic stages

Group 3:

Plasmodium berghei 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

Samples:

Erythrocytic stages (Asexual):
SAMN02428730
SAMN02428734

Salivary gland sporozoites:
SAMN02428726
SAMN02428729

Cultured sporozoites:
SAMN02428728
SAMN02428727

Group 4:

Plasmodium falciparum salivary sporozoites vs. erythrocytic stages

Group 5:

Plasmodium falciparum cultured sporozoites vs. erythrocytic stages

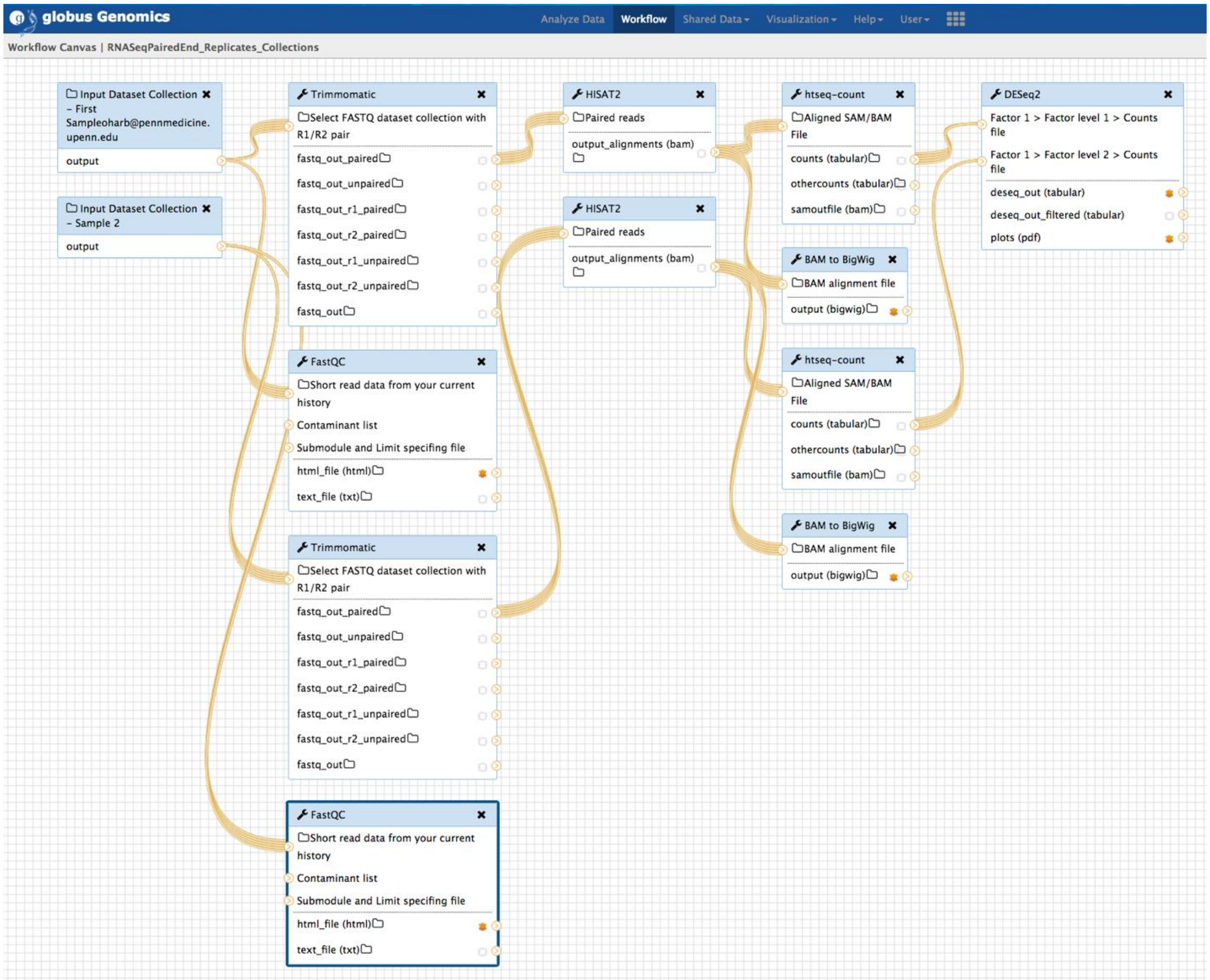
Group 6:

Plasmodium falciparum salivary sporozoites vs. cultured sporozoites

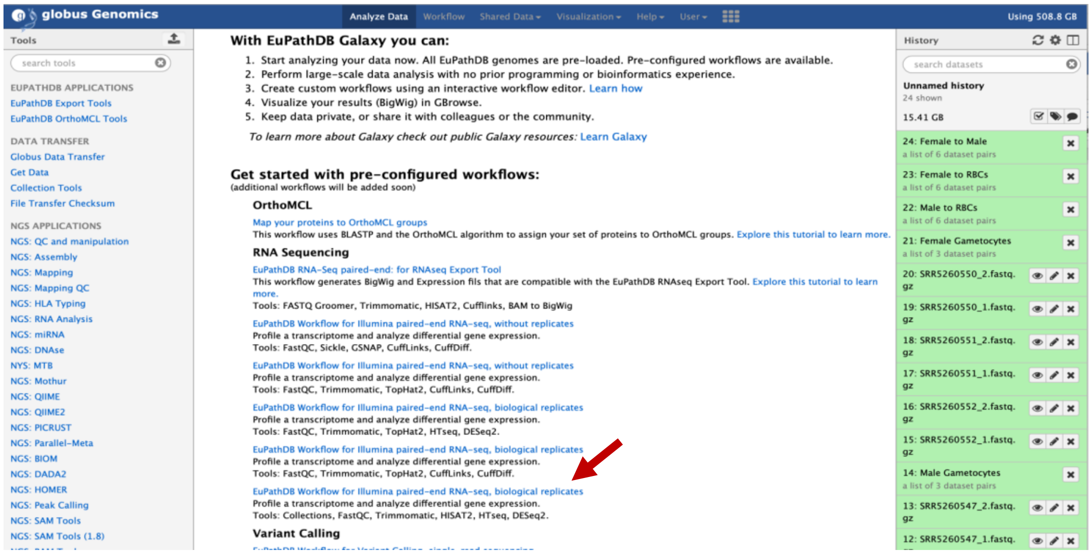
Section II: Running a workflow in Galaxy

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



EUPATHDB APPLICATIONS
EuPathDB Export Tools
EuPathDB OrthoMCL Tools

DATA TRANSFER
Globus Data Transfer
Get Data
Collection Tools
File Transfer Checksum

NGS APPLICATIONS
NGS: QC and manipulation
NGS: Assembly
NGS: Mapping
NGS: Mapping QC
NGS: HLA Typing
NGS: RNA Analysis
NGS: miRNA
NGS: DNAse
NYS: MTB
NGS: Mothur
NGS: QIIME
NGS: QIIME2
NGS: PICRUST
NGS: Parallel-Meta
NGS: BIOM
NGS: DADA2
NGS: HOMER
NGS: Peak Calling
NGS: SAM Tools
NGS: SAM Tools (1.8)

With EuPathDB Galaxy you can:

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

To learn more about Galaxy check out public Galaxy resources: Learn Galaxy

Get started with pre-configured workflows:
(additional workflows will be added soon)

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 fils that are compatible with the EuPathDB RNAseq Export Tool. Explore this tutorial to learn more.
Tools: FASTQ Groomer, Trimmomatic, HISAT2, Cufflinks, BAM to BigWig

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.

EuPathDB Workflow for Illumina paired-end RNA-seq, biological 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: Collections, FastQC, Trimmomatic, HISAT2, HTseq, DESeq2.

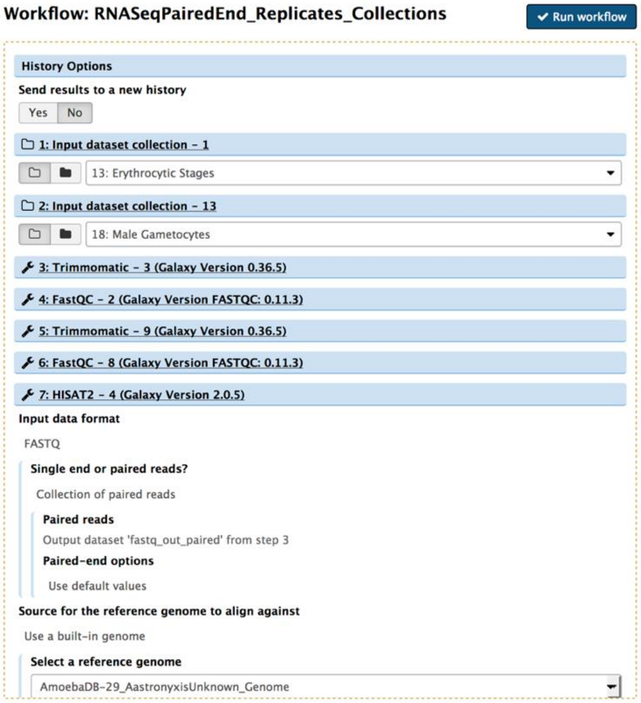
Variant Calling

Unnamed history
24 shown
15.41 GB

24: Female to Male
a list of 6 dataset pairs
23: Female to RBCs
a list of 6 dataset pairs
22: Male to RBCs
a list of 6 dataset pairs
21: Female Gametocytes
a list of 3 dataset pairs
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
a list of 3 dataset pairs
13: SRR5260547_2.fastq.gz
12: SRR5260547_1.fastq.

- 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.



Workflow: RNASeqPairedEnd_Replicates_Collections

History Options
Send results to a new history
Yes No

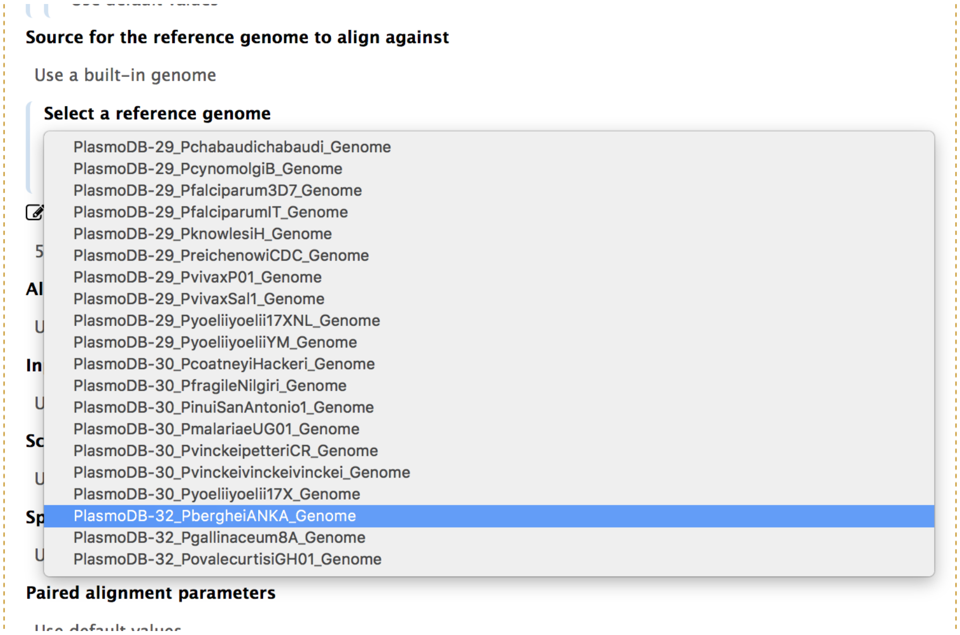
1: Input dataset collection - 1
13: Erythrocytic Stages

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_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

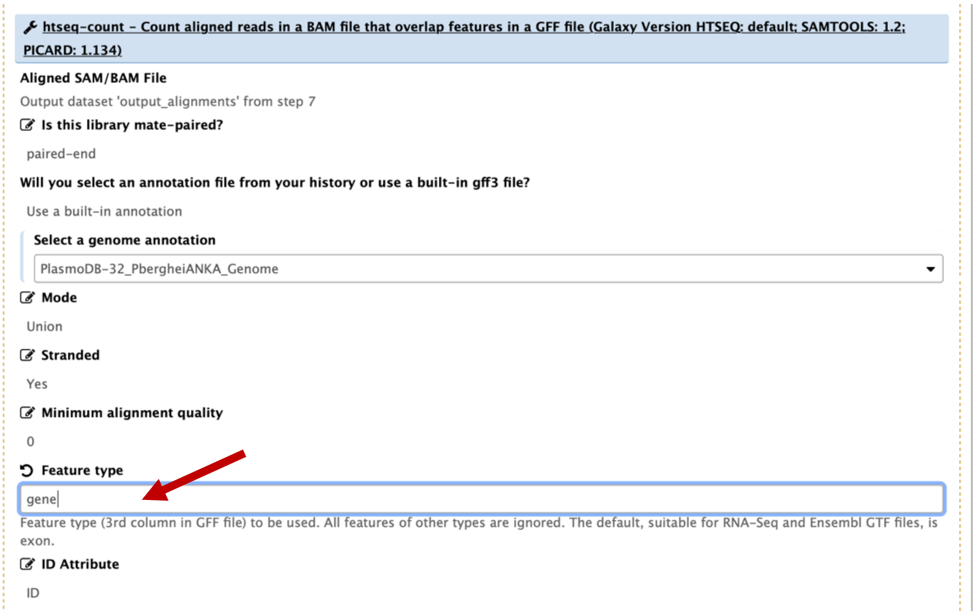
Use a built-in genome

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_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_PbergheiANKA_Genome
PlasmoDB-32_Pgallinaceum8A_Genome
PlasmoDB-32_PovalecurtisiGH01_Genome

Paired alignment parameters

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 chance 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

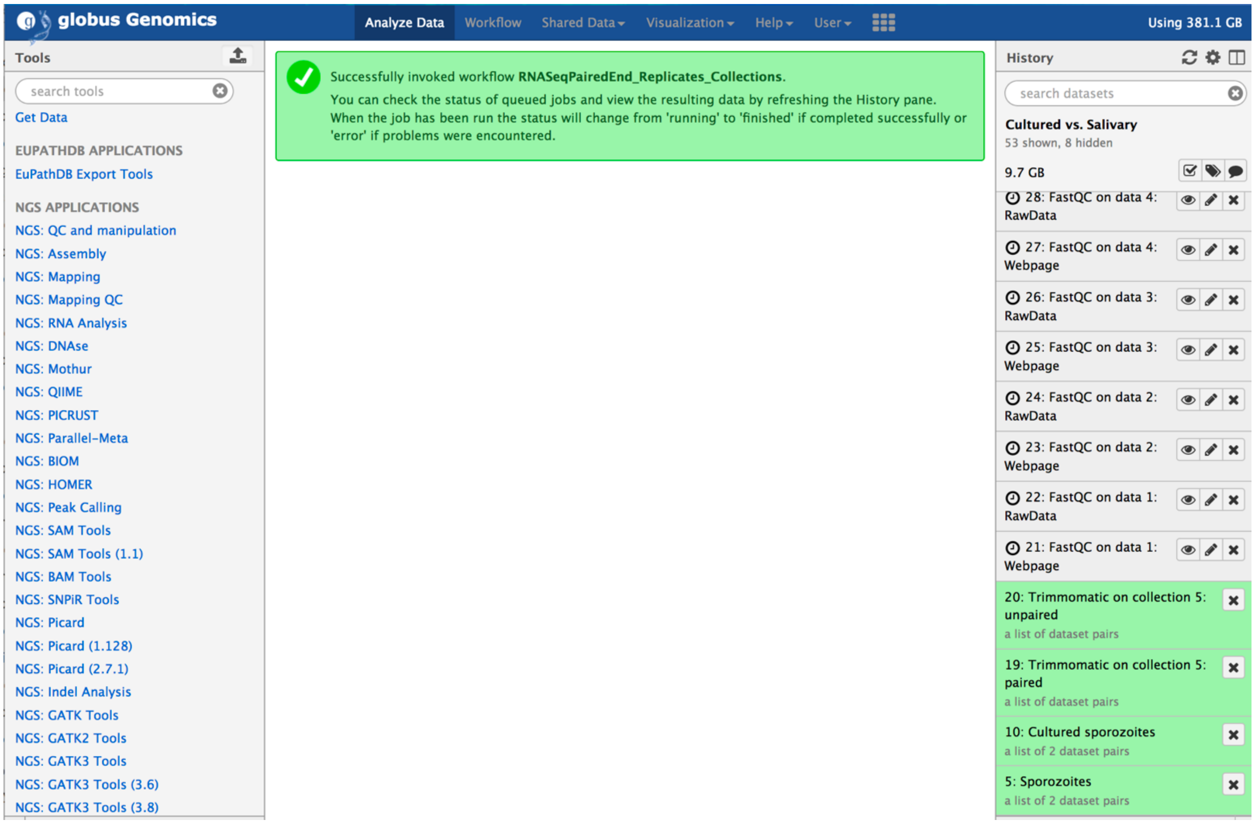
gene

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.



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

search datasets

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
a list of dataset pairs

19: Trimmomatic on collection 5: paired
a list of dataset pairs

10: Cultured sporozoites
a list of 2 dataset pairs

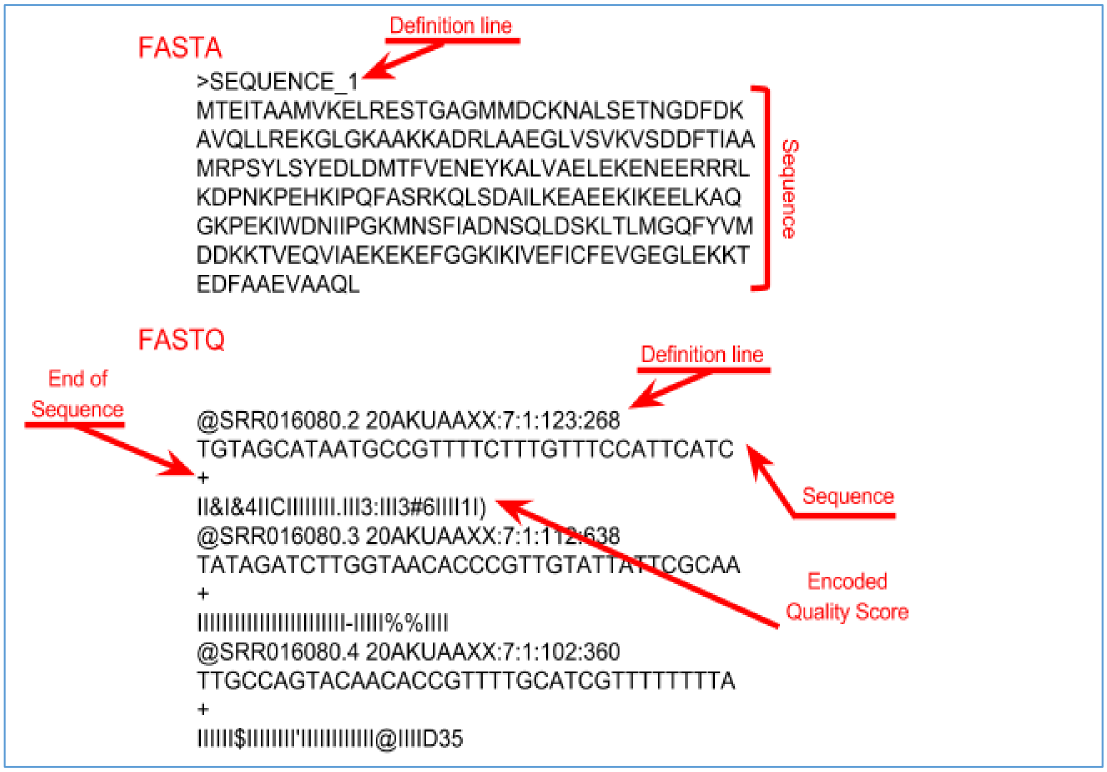
5: Sporozoites
a list of 2 dataset pairs

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


FASTA
Definition line
Sequence
FASTQ
Definition line
End of Sequence
Sequence
Encoded Quality Score

```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDK
AVQLLREKGLGKAAKKADRLAAEGLVSVKVSDDFTIAA
MRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRL
KDPNKPEHKIPQFASRKQLSDAILKEAEEKIKEELKAQ
GKPEKIWDNIIPGKMNSFIADNSQLDSKLTLMGQFYVM
DDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKT
EDFAAEVAAQL
```

```
@SRR016080.2 20AKUAAXX:7:1:123:268
TGTAGCATAATGCCGTTTTCTTTGTTTCCATTCATC
+
II&I&4IICIIIIII.III3:III3#6IIIII1I)
@SRR016080.3 20AKUAAXX:7:1:112:638
TATAGATCTTGGTAACACCCGTTGTATTATTCGCAA
+
IIIIIIIIIIIIIIIIIIIIIIII-IIIII%%IIII
@SRR016080.4 20AKUAAXX:7:1:102:360
TTGCCAGTACAACACCGTTTTGCATCGTTTTTTTTA
+
IIIIII$IIIIIIII'IIIIIIIIIIII@IIIID35
```