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

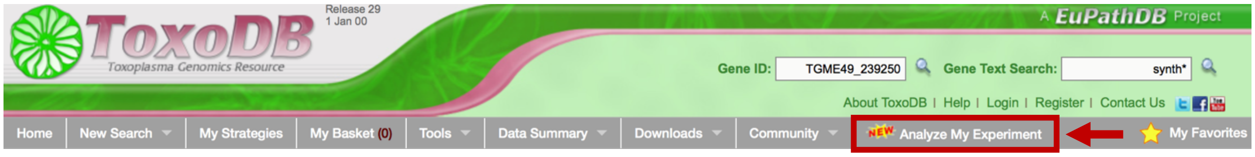
The goal of this exercise is to use a Galaxy workflow to analyze RNA sequencing data. The datasets we will us are all feely available through the sequence repositories (SRA, ENA, DDBJ).

Galaxy is an open, web-based platform for data intensive biomedical research. EuPathDB is developing its own Galaxy instance that will become available for its users. Galaxy allows you to perform, reproduce, and share complete analyses.

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

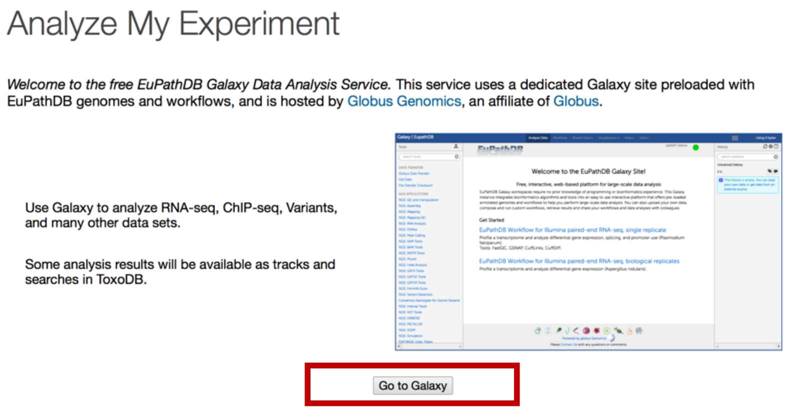
For this exercise we will be working in groups. Each group will have 4-6 members. One person in the group will run the Galaxy controls on one computer. The other members' role is to help ensure that the correct datasets are being used and that the correct workflow parameters are being selected. You can access the EuPathDB Galaxy instance through the "Analyze My Experiment" link in the gray menu bar in any EuPathDB site:



Gene ID: TGME49_239250
Gene Text Search: synth*

Section I: Setting up your EuPathDB Galaxy account

Step 1: Click on the "Analyze My Experiment" link in the gray menu bar.
Step 2: On the next page you will see a description of this service. In order to start using the EuPathDB Galaxy instance you will have to follow the registration steps. Start by clicking on the "Continue with Galaxy Sign-up" button.



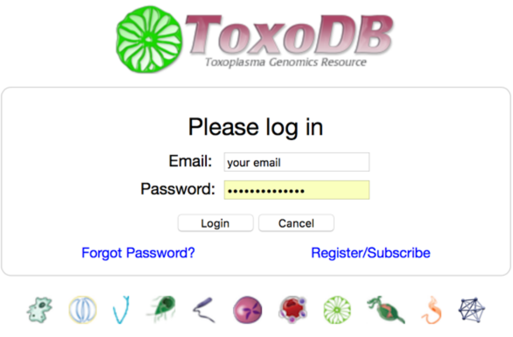
Analyze My Experiment

Welcome to the free EuPathDB Galaxy Data Analysis Service. This service uses a dedicated Galaxy site preloaded with EuPathDB genomes and workflows, and is hosted by Globus Genomics, an affiliate of Globus.

Use Galaxy to analyze RNA-seq, ChIP-seq, Variants, and many other data sets.

Some analysis results will be available as tracks and searches in ToxoDB.

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



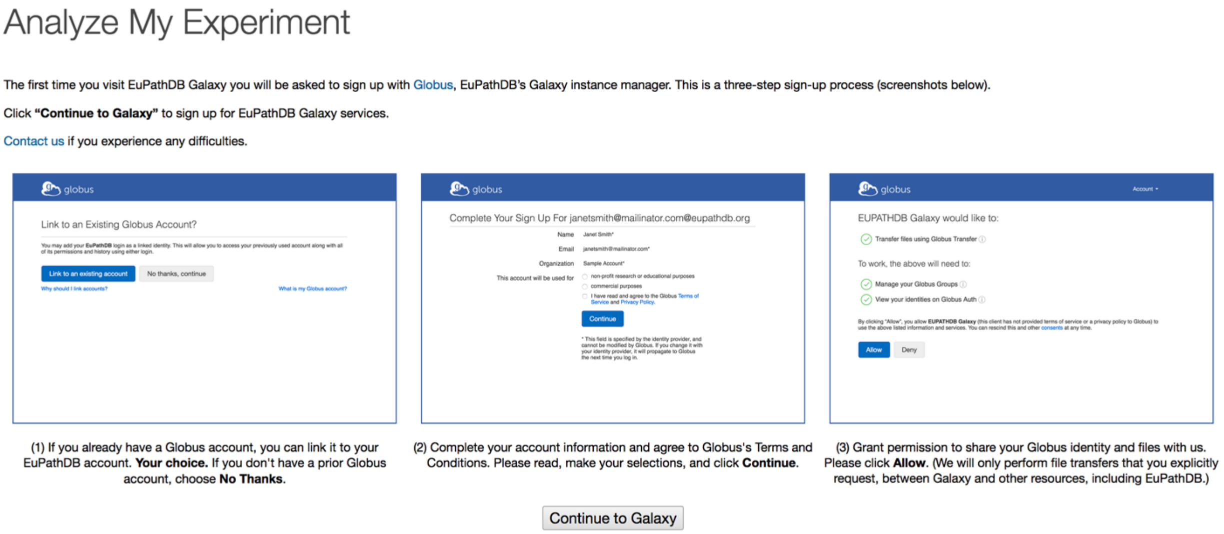
ToxoDB
Toxoplasma Genomics Resource

Please log in

Email: your email

Password: ••••••••••••••

Step 4: The next window describes the process of signing 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.

globus

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.

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

(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.**

globus

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.

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

globus

EUPATHDB Galaxy would like to:

✓ Transfer files using Globus Transfer ⓘ

To work, the above will need to:

✓ Manage your Globus Groups ⓘ

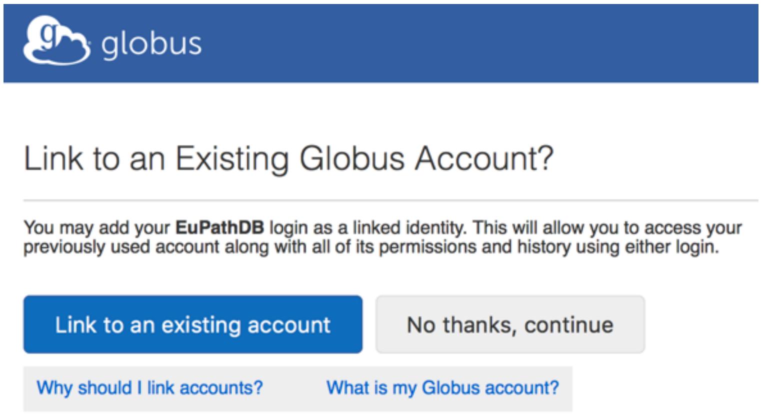
✓ View your identities on Globus Auth ⓘ

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.

(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"



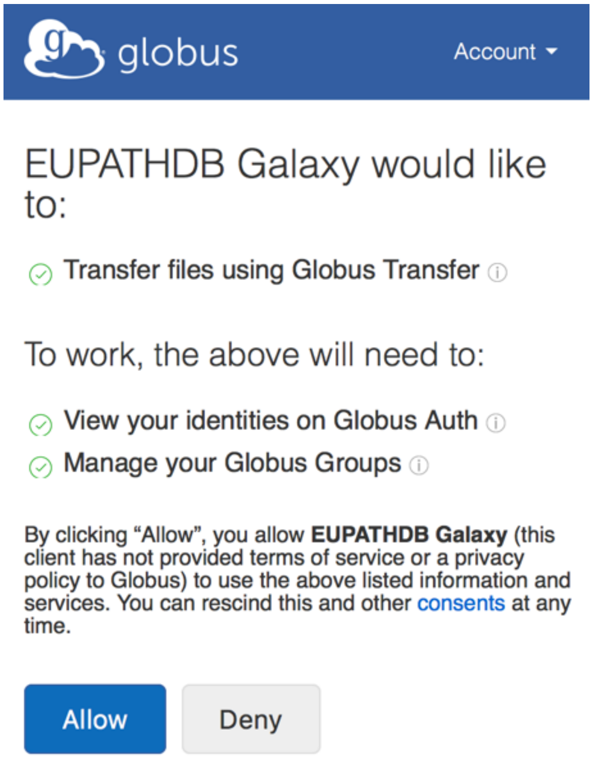
globus

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.

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 ⓘ

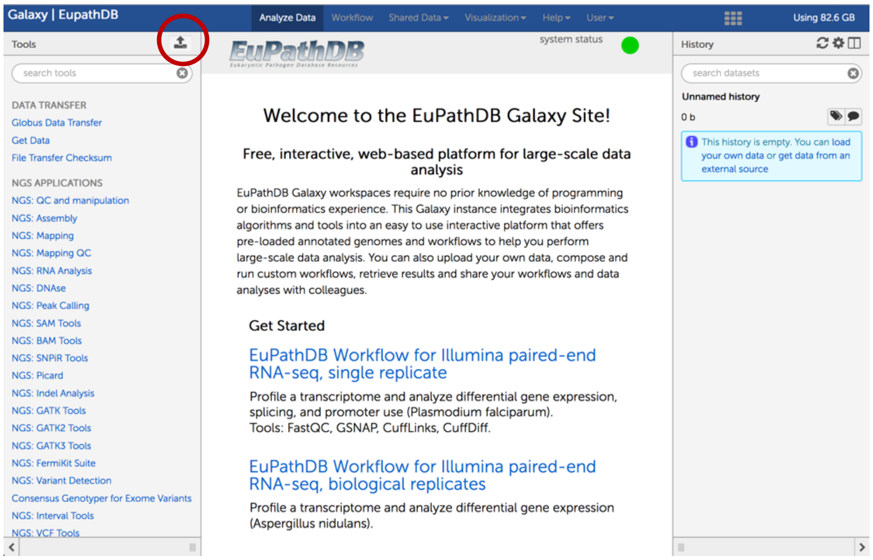
By clicking “Allow”, you allow **EUPATHDB Galaxy** (this client has not provided terms of service or a privacy policy to Globus) to use the above listed information and services. You can rescind this and other consents at any time.

Step 8: Congratulations, you are in!

<u>Section II: Importing data to Galaxy</u>

There are multiple ways to important data into your Galaxy workspace. For this exercise we will use the direct links listed below. Remember one person in your group will be doing this. 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” icon. This will open up a window that allows you to “**Download from web or upload from disk”**



This history is empty. You can load your own data or get data from an external source

Welcome to the EuPathDB Galaxy Site!

Free, interactive, web-based platform for large-scale data analysis

EuPathDB Galaxy workspaces require no prior knowledge of programming or bioinformatics experience. This Galaxy instance integrates bioinformatics algorithms and tools into an easy to use interactive platform that offers pre-loaded annotated genomes and workflows to help you perform large-scale data analysis. You can also upload your own data, compose and run custom workflows, retrieve results and share your workflows and data analyses with colleagues.

Get Started

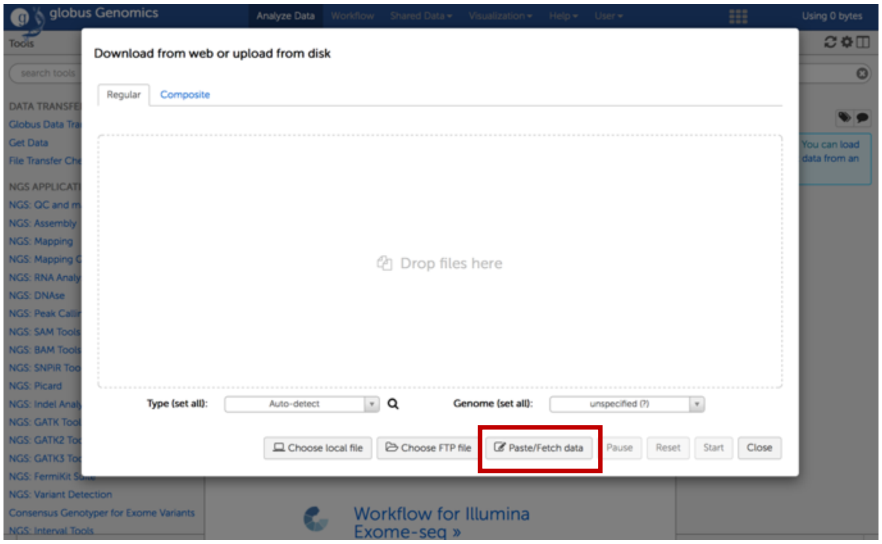
EuPathDB Workflow for Illumina paired-end RNA-seq, single replicate

Profile a transcriptome and analyze differential gene expression, splicing, and promoter use (Plasmodium falciparum). Tools: FastQC, GSNAP, CuffLinks, CuffDiff.

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

Profile a transcriptome and analyze differential gene expression (Aspergillus nidulans).

Step 2: In the “Download from web or upload from disk” window click on “Paste/Fetch data”

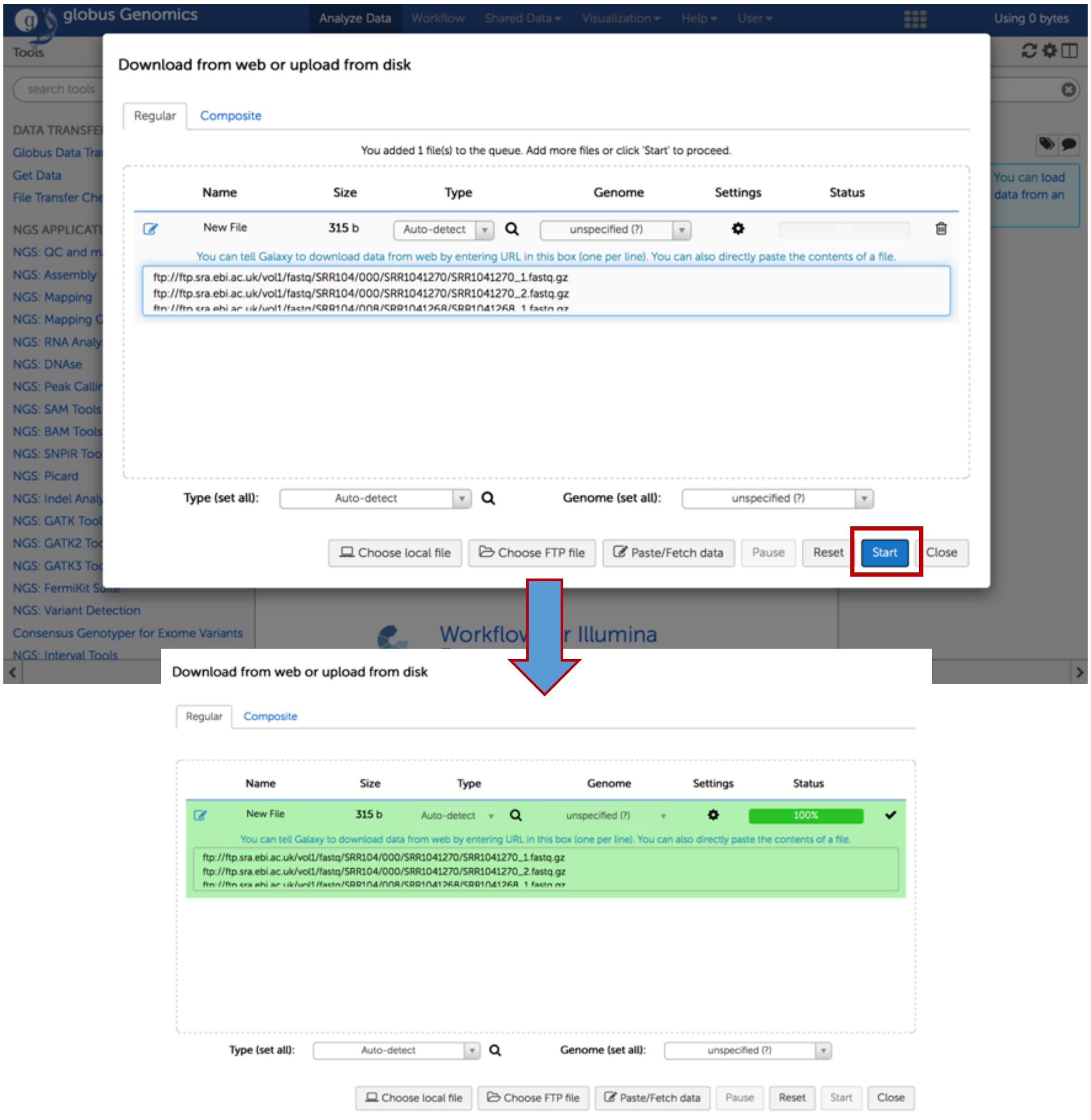


Download from web or upload from disk

Type (set all): Auto-detect

Genome (set all): unspecified (?)

Step 3: Paste the four URLs corresponding to the four files for your group. Each URL has to be on a new line. Then click on "Start".



Download from web or upload from disk

You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.

Name	Size	Type	Genome	Settings	Status
New File	315 b	Auto-detect	unspecified (?)		

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

```
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_2.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_1.fastq.gz
```

Type (set all): Auto-detect
Genome (set all): unspecified (?)

Download from web or upload from disk

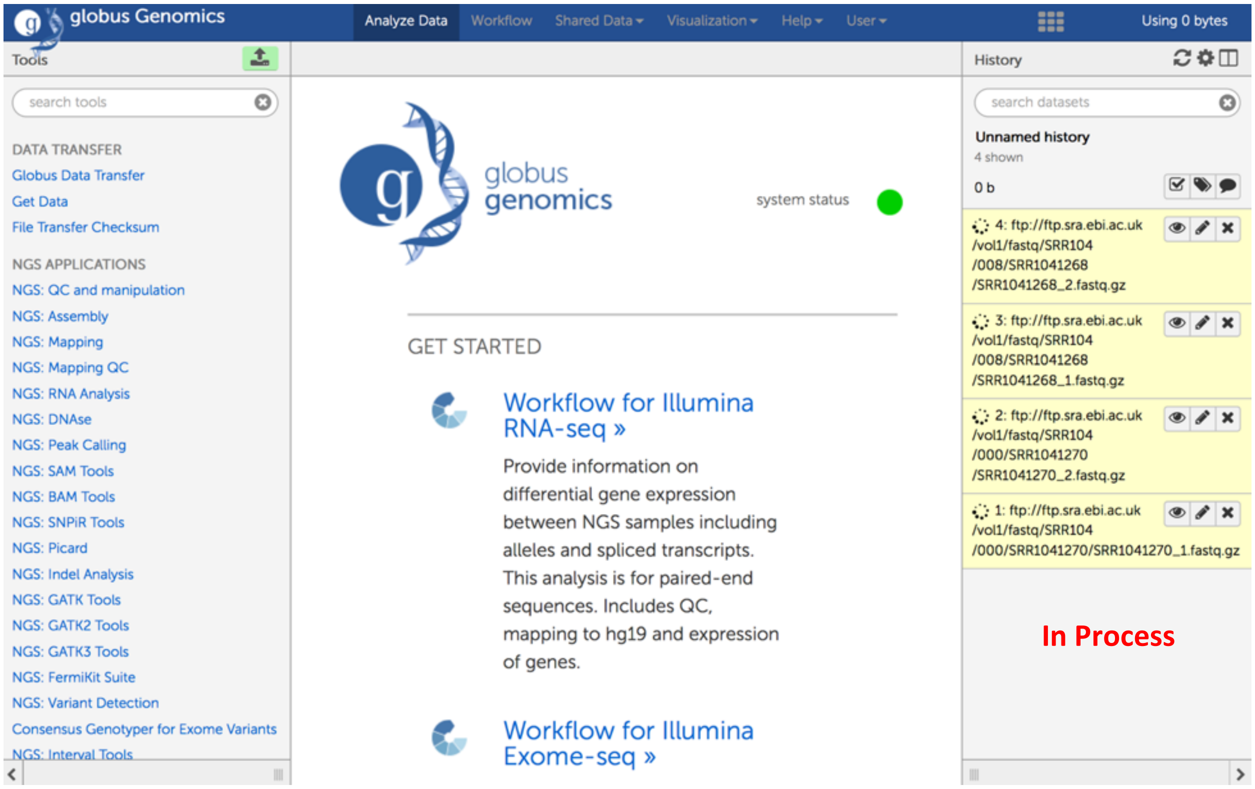
Name	Size	Type	Genome	Settings	Status
New File	315 b	Auto-detect	unspecified (?)		100% ✓

You can tell Galaxy to download data from web by entering URL in this box (one per line). You can also directly paste the contents of a file.

```
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_2.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_1.fastq.gz
```

Type (set all): Auto-detect
Genome (set all): unspecified (?)

Step 4: Click on "Close". You should notice that the left section (history section) will show the files being transferred (yellow) – this may take a few minutes to start. File transfer will take about 15-20 minutes. When this is complete they will turn green.



DATA TRANSFER
Globus Data Transfer
Get Data
File Transfer Checksum

NGS APPLICATIONS
NGS: QC and manipulation
NGS: Assembly
NGS: Mapping
NGS: Mapping QC
NGS: RNA Analysis
NGS: DNAse
NGS: Peak Calling
NGS: SAM Tools
NGS: BAM Tools
NGS: SNPiR Tools
NGS: Picard
NGS: Indel Analysis
NGS: GATK Tools
NGS: GATK2 Tools
NGS: GATK3 Tools
NGS: FermiKit Suite
NGS: Variant Detection
Consensus Genotyper for Exome Variants
NGS: Interval Tools

globus genomics

GET STARTED

Workflow for Illumina RNA-seq »

Provide information on differential gene expression between NGS samples including alleles and spliced transcripts. This analysis is for paired-end sequences. Includes QC, mapping to hg19 and expression of genes.

Workflow for Illumina Exome-seq »

Unnamed history
4 shown
0 b

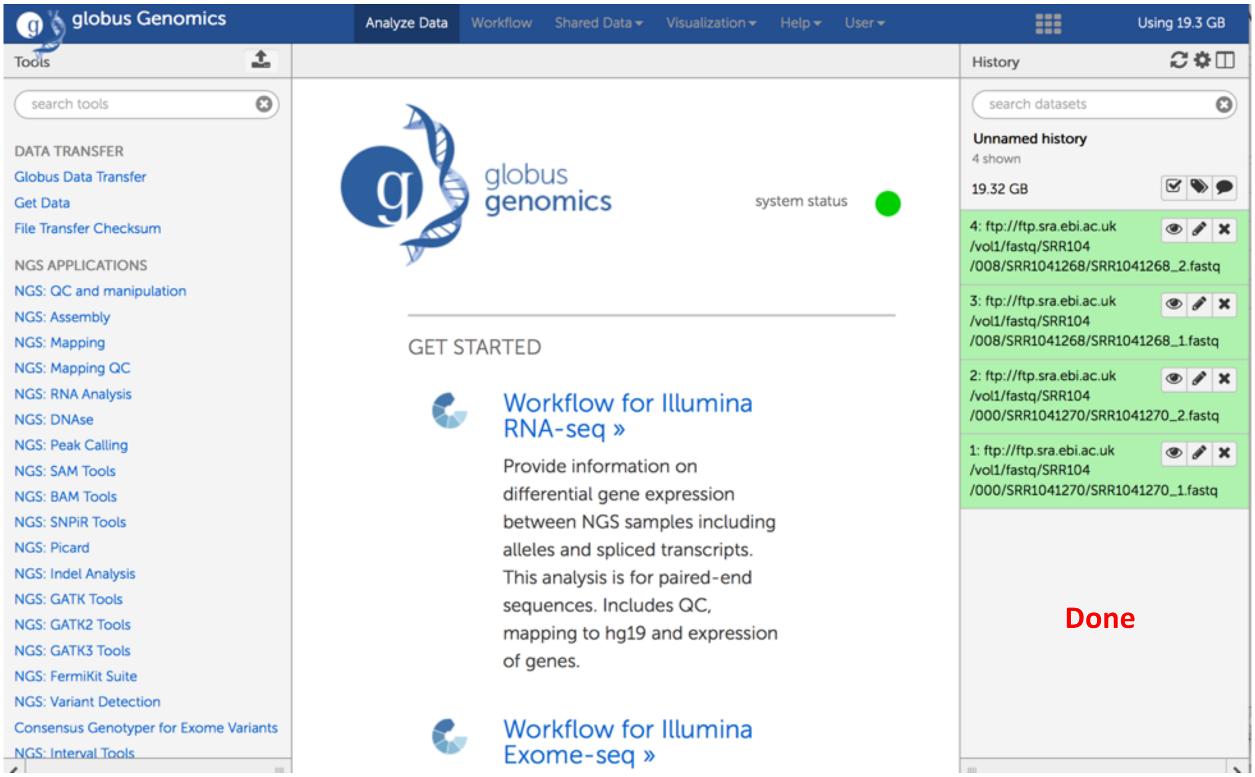
4: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_2.fastq.gz

3: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_1.fastq.gz

2: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_2.fastq.gz

1: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_1.fastq.gz

In Process



DATA TRANSFER
Globus Data Transfer
Get Data
File Transfer Checksum

NGS APPLICATIONS
NGS: QC and manipulation
NGS: Assembly
NGS: Mapping
NGS: Mapping QC
NGS: RNA Analysis
NGS: DNAse
NGS: Peak Calling
NGS: SAM Tools
NGS: BAM Tools
NGS: SNPiR Tools
NGS: Picard
NGS: Indel Analysis
NGS: GATK Tools
NGS: GATK2 Tools
NGS: GATK3 Tools
NGS: FermiKit Suite
NGS: Variant Detection
Consensus Genotyper for Exome Variants
NGS: Interval Tools

globus genomics

GET STARTED

Workflow for Illumina RNA-seq »

Provide information on differential gene expression between NGS samples including alleles and spliced transcripts. This analysis is for paired-end sequences. Includes QC, mapping to hg19 and expression of genes.

Workflow for Illumina Exome-seq »

Unnamed history
4 shown
19.32 GB

4: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_2.fastq

3: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_1.fastq

2: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_2.fastq

1: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_1.fastq

Done

Group assignments:

Group 1:

Plasmodium falciparum Asexual vs. Cultured sporozoites
Project information: http://www.ebi.ac.uk/ena/data/view/PRJNA230379

Samples:

Asexual samples:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_2.fastq.gz

Cultured sporozoite samples:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_2.fastq.gz

Group 2:

Plasmodium falciparum Asexual vs. Salivary sporozoites
Project information: http://www.ebi.ac.uk/ena/data/view/PRJNA230379

Samples:

Asexual samples:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_2.fastq.gz

Salivary Sporozoites:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/006/SRR1041266/SRR1041266_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/006/SRR1041266/SRR1041266_2.fastq.gz

Group 3:

Plasmodium falciparum Cultured vs. Salivary sporozoites
Project information: http://www.ebi.ac.uk/ena/data/view/PRJNA230379

Samples:

Cultured sporozoite samples:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_2.fastq.gz

Salivary Sporozoites:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/006/SRR1041266/SRR1041266_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/006/SRR1041266/SRR1041266_2.fastq.gz

Group 4:

Aspergillus nidulans FGSC4 VeA$^+$ WT vs. OSA knock outs
Project information: http://www.ebi.ac.uk/ena/data/view/PRJNA293709

Samples:

FGSC4 VeA$^+$ WT:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR218/001/SRR2180251/SRR2180251_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR218/001/SRR2180251/SRR2180251_2.fastq.gz

OSA knock outs:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR218/007/SRR2180257/SRR2180257_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR218/007/SRR2180257/SRR2180257_2.fastq.gz

Group 5:

Toxoplasma gondii WT vs. GRA17 knock outs
Project information: http://www.ebi.ac.uk/ena/data/view/PRJNA275621

Samples:

WT:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR180/001/SRR1805881/SRR1805881_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR180/001/SRR1805881/SRR1805881_2.fastq.gz

GRA17 knock outs:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR180/002/SRR1805882/SRR1805882_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR180/002/SRR1805882/SRR1805882_2.fastq.gz

Group 6:

Toxoplasma gondii WT vs. GRA17 knock outs
Project information: http://www.ebi.ac.uk/ena/data/view/PRJNA275621

Samples:

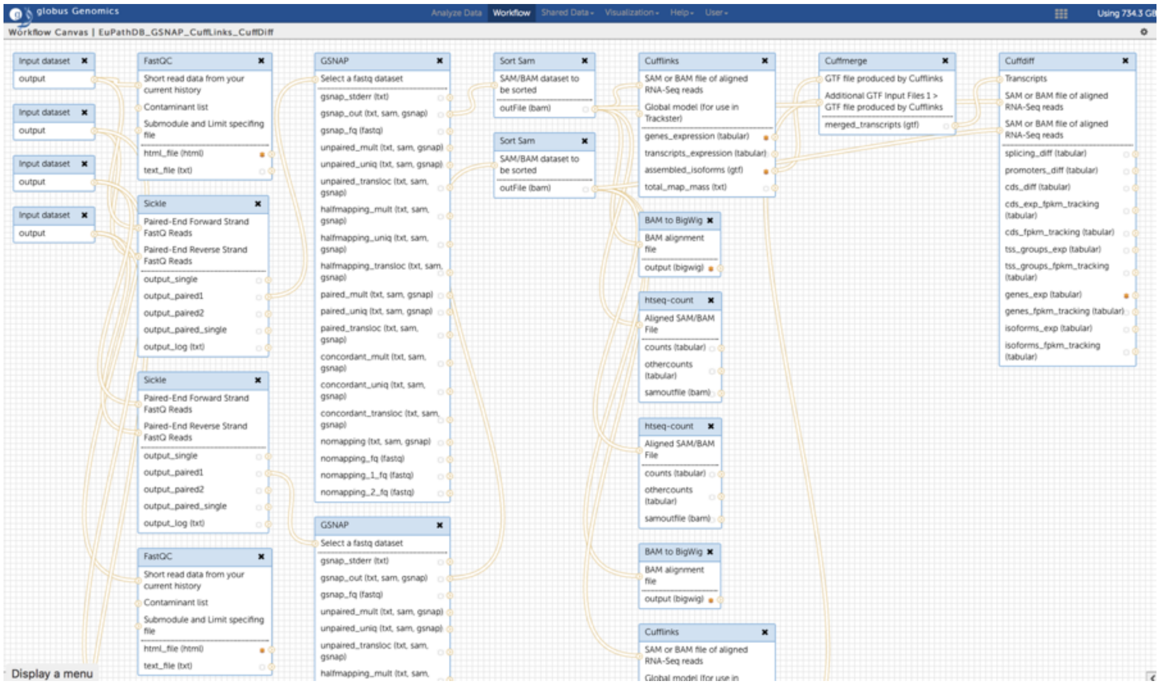
WT:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR180/001/SRR1805881/SRR1805881_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR180/001/SRR1805881/SRR1805881_2.fastq.gz

GRA23 knock outs:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR180/003/SRR1805883/SRR1805883_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR180/003/SRR1805883/SRR1805883_2.fastq.gz

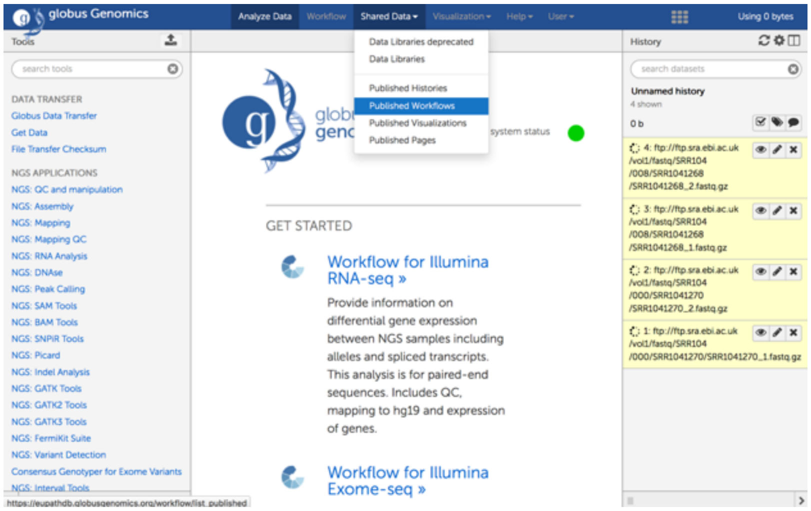
Section II: Running a workflow in Galaxy

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.
3. Aligns the reads to a reference genome using GSNAP and generates coverage plots.
4. Determines FPKM values for each sample and generates gene/transcript models.
5. Determines differential expression of genes between the samples.

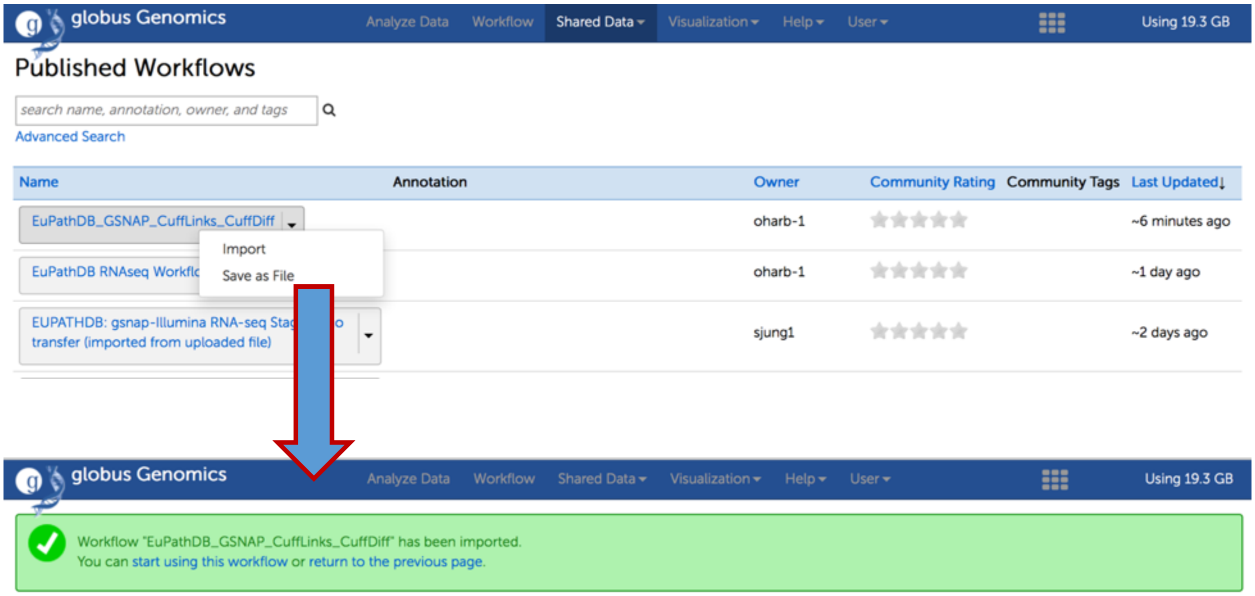

globus Genomics
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 734.3 GB
Workflow Canvas | EuPathDB_GSNAP_CuffLinks_CuffDiff
Input dataset
output
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
Sickle
Paired-End Forward Strand FastQ Reads
Paired-End Reverse Strand FastQ Reads
output_single
output_paired1
output_paired2
output_paired_single
output_log (txt)
GSNAP
Select a fastq dataset
gsnap_stderr (txt)
gsnap_out (txt, sam, gsnap)
gsnap_fq (fastq)
unpaired_mult (txt, sam, gsnap)
unpaired_uniq (txt, sam, gsnap)
unpaired_transloc (txt, sam, gsnap)
halfmapping_mult (txt, sam, gsnap)
halfmapping_uniq (txt, sam, gsnap)
halfmapping_transloc (txt, sam, gsnap)
paired_mult (txt, sam, gsnap)
paired_uniq (txt, sam, gsnap)
paired_transloc (txt, sam, gsnap)
concordant_mult (txt, sam, gsnap)
concordant_uniq (txt, sam, gsnap)
concordant_transloc (txt, sam, gsnap)
nomapping (txt, sam, gsnap)
nomapping_fq (fastq)
nomapping_1_fq (fastq)
nomapping_2_fq (fastq)
Sort Sam
SAM/BAM dataset to be sorted
outFile (bam)
Cufflinks
SAM or BAM file of aligned RNA-Seq reads
Global model (for use in Trackster)
genes_expression (tabular)
transcripts_expression (tabular)
assembled_isoforms (gtf)
total_map_mass (txt)
BAM to BigWig
BAM alignment file
output (bigwig)
htseq-count
Aligned SAM/BAM File
counts (tabular)
othercounts (tabular)
samoutfile (bam)
Cuffmerge
GTF file produced by Cufflinks
Additional GTF Input Files 1 > GTF file produced by Cufflinks
merged_transcripts (gtf)
Cuffdiff
Transcripts
SAM or BAM file of aligned RNA-Seq reads
splicing_diff (tabular)
promoters_diff (tabular)
cds_diff (tabular)
cds_exp_fpkm_tracking (tabular)
cds_fpkm_tracking (tabular)
tss_groups_exp (tabular)
tss_groups_fpkm_tracking (tabular)
genes_exp (tabular)
genes_fpkm_tracking (tabular)
isoforms_exp (tabular)
isoforms_fpkm_tracking (tabular)
Display a menu

Step 1: Import the workflow called "EuPathDB_GSNAP_CuffLinks_CuffDiff" – click on the shared data menu item and select "Published Workflows" from the menu.



Provide information on differential gene expression between NGS samples including alleles and spliced transcripts. This analysis is for paired-end sequences. Includes QC, mapping to hg19 and expression of genes.

Step 2: Click on the arrow next to the appropriate workflow and select import.



Published Workflows

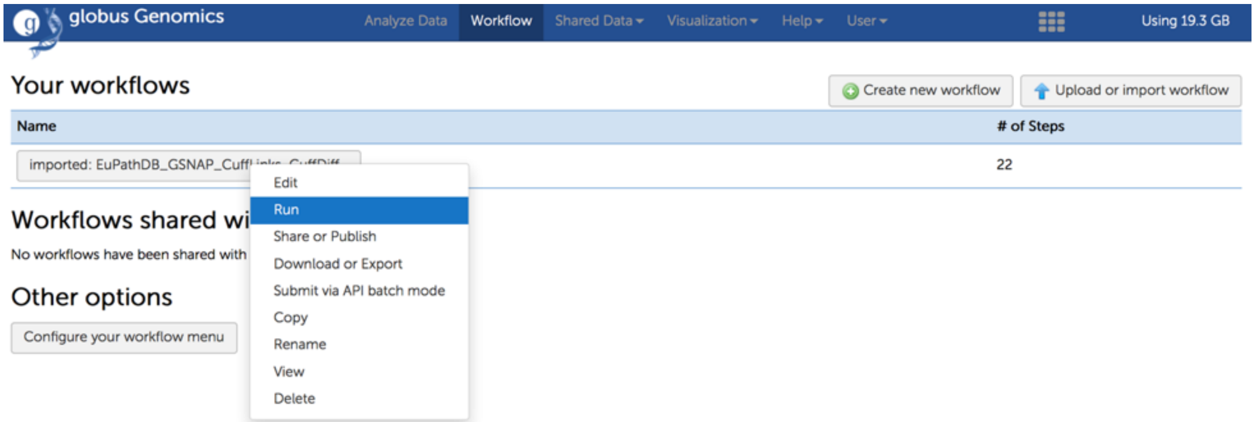
search name, annotation, owner, and tags

Advanced Search

Name	Annotation	Owner	Community Rating	Community Tags	Last Updated↓
EuPathDB_GSNAP_CuffLinks_CuffDiff		oharb-1	☆☆☆☆☆		~6 minutes ago
EuPathDB RNAseq Workflo		oharb-1	☆☆☆☆☆		~1 day ago
EUPATHDB: gsnap-Illumina RNA-seq Stag o transfer (imported from uploaded file)		sjung1	☆☆☆☆☆		~2 days ago

Workflow "EuPathDB_GSNAP_CuffLinks_CuffDiff" has been imported.
You can start using this workflow or return to the previous page.

Step 3: Click on "Workflow" in the menu at the top of the page. On the next page click on the arrow next to your imported workflow and select the "Run" option.



Your workflows

Name	# of Steps
imported: EuPathDB_GSNAP_Cuff	22

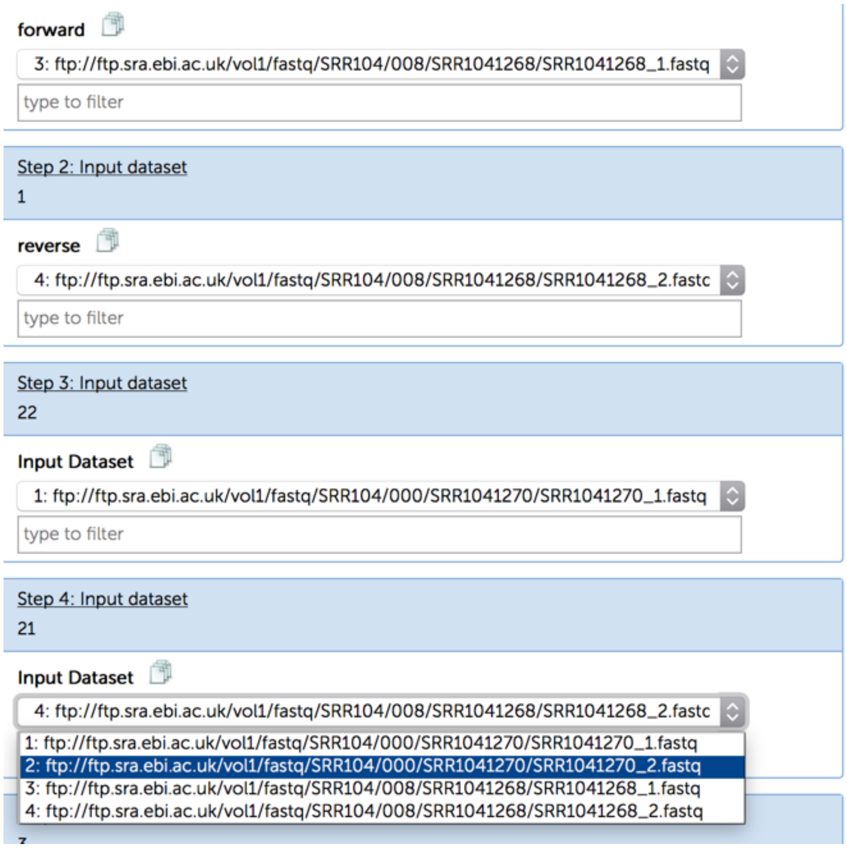
Workflows shared wi

No workflows have been shared with

Other options

Step 4: 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 datasets. These are the fastq files you imported from the sequence archive. Workflow steps 1-4 allow you to select the datasets. Be sure you match the correct forward and reverse files. The should end in the same SRR number with a .1 or .2 at the end.



forward

3: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_1.fastq

type to filter

Step 2: Input dataset

1

reverse

4: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_2.fastc

type to filter

Step 3: Input dataset

22

Input Dataset

1: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_1.fastq

type to filter

Step 4: Input dataset

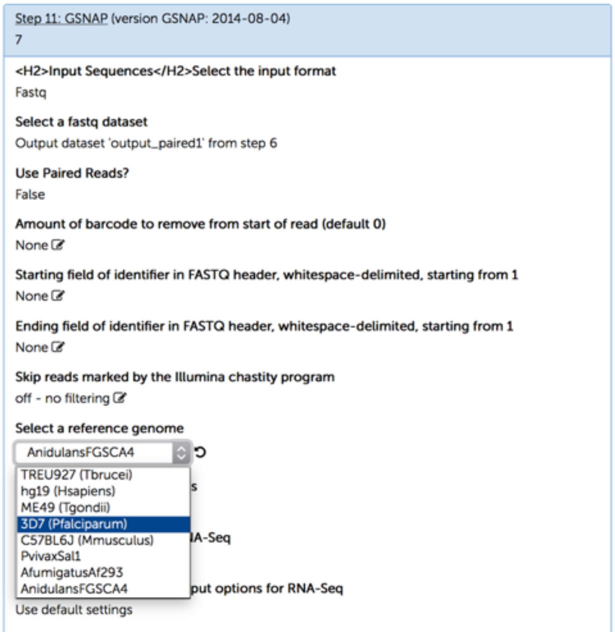
21

Input Dataset

4: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_2.fastc

1: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_1.fastq
2: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_2.fastq
3: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_1.fastq
4: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_2.fastq

b. Scroll down to steps 11 and 12 (GSNAP). Click on the name of the step to open up the parameters. Select the correct reference organism in each of the steps.



Step 11: GSNAP (version GSNAP: 2014-08-04)
7

<H2>Input Sequences</H2>Select the input format
Fastq

Select a fastq dataset
Output dataset 'output_paired1' from step 6

Use Paired Reads?
False

Amount of barcode to remove from start of read (default 0)
None

Starting field of identifier in FASTQ header, whitespace-delimited, starting from 1
None

Ending field of identifier in FASTQ header, whitespace-delimited, starting from 1
None

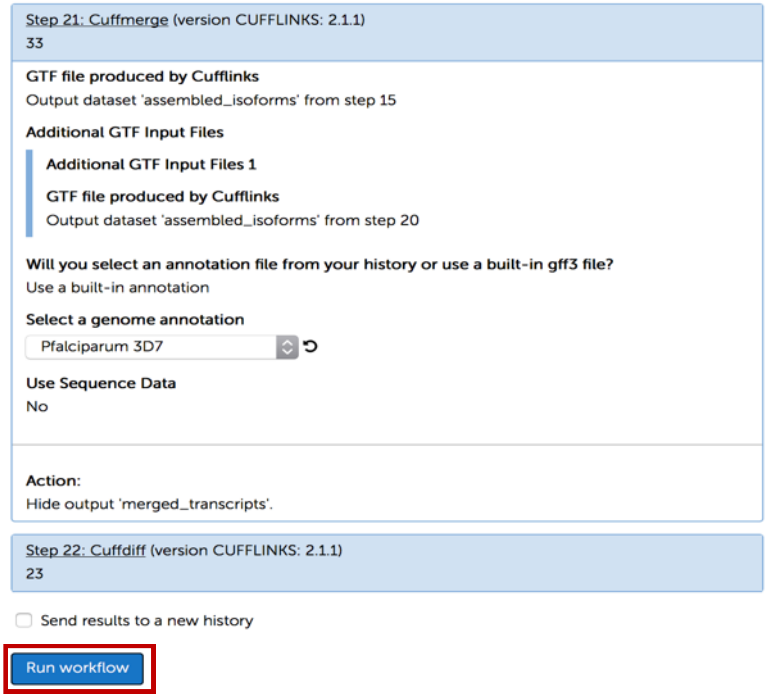
Skip reads marked by the Illumina chastity program
off - no filtering

Select a reference genome
AnidulansFGSCA4

Use default settings

c. Scroll down to step 15 (Cufflinks), 17, 18 (htseq), 20 (Cufflinks) and 21 (Cuffmerge) and select the correct reference organism.

d. Click on "Run Workflow"



Step 21: Cuffmerge (version CUFFLINKS: 2.1.1)
33

GTF file produced by Cufflinks
Output dataset 'assembled_isoforms' from step 15

Additional GTF Input Files

Additional GTF Input Files 1

GTF file produced by Cufflinks
Output dataset 'assembled_isoforms' from step 20

Will you select an annotation file from your history or use a built-in gff3 file?
Use a built-in annotation

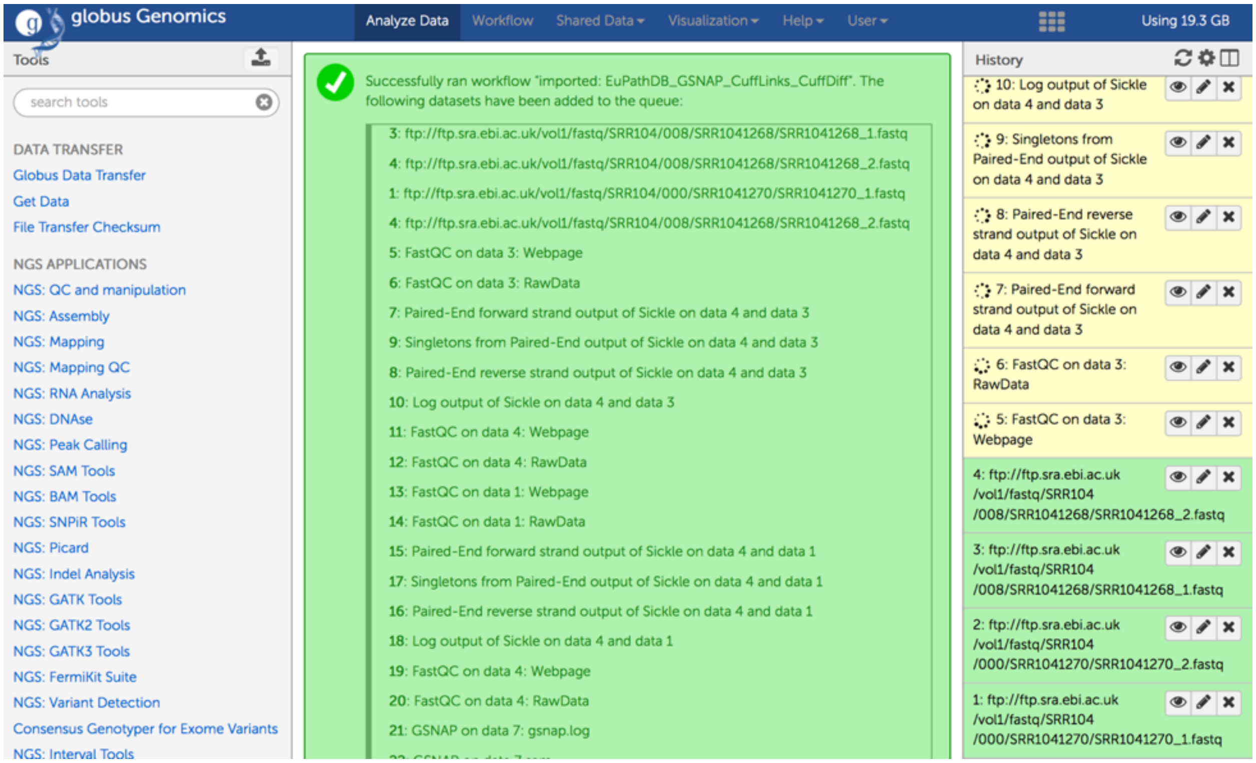
Select a genome annotation
Pfalciparum 3D7

Use Sequence Data
No

Action:
Hide output 'merged_transcripts'.

Step 22: Cuffdiff (version CUFFLINKS: 2.1.1)
23

☐ Send results to a new history



Successfully ran workflow "imported: EuPathDB_GSNAP_CuffLinks_CuffDiff". The following datasets have been added to the queue:

3: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_1.fastq
4: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_2.fastq
1: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/000/SRR1041270/SRR1041270_1.fastq
4: ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR104/008/SRR1041268/SRR1041268_2.fastq
5: FastQC on data 3: Webpage
6: FastQC on data 3: RawData
7: Paired-End forward strand output of Sickle on data 4 and data 3
9: Singletons from Paired-End output of Sickle on data 4 and data 3
8: Paired-End reverse strand output of Sickle on data 4 and data 3
10: Log output of Sickle on data 4 and data 3
11: FastQC on data 4: Webpage
12: FastQC on data 4: RawData
13: FastQC on data 1: Webpage
14: FastQC on data 1: RawData
15: Paired-End forward strand output of Sickle on data 4 and data 1
17: Singletons from Paired-End output of Sickle on data 4 and data 1
16: Paired-End reverse strand output of Sickle on data 4 and data 1
18: Log output of Sickle on data 4 and data 1
19: FastQC on data 4: Webpage
20: FastQC on data 4: RawData
21: GSNAP on data 7: gsnap.log

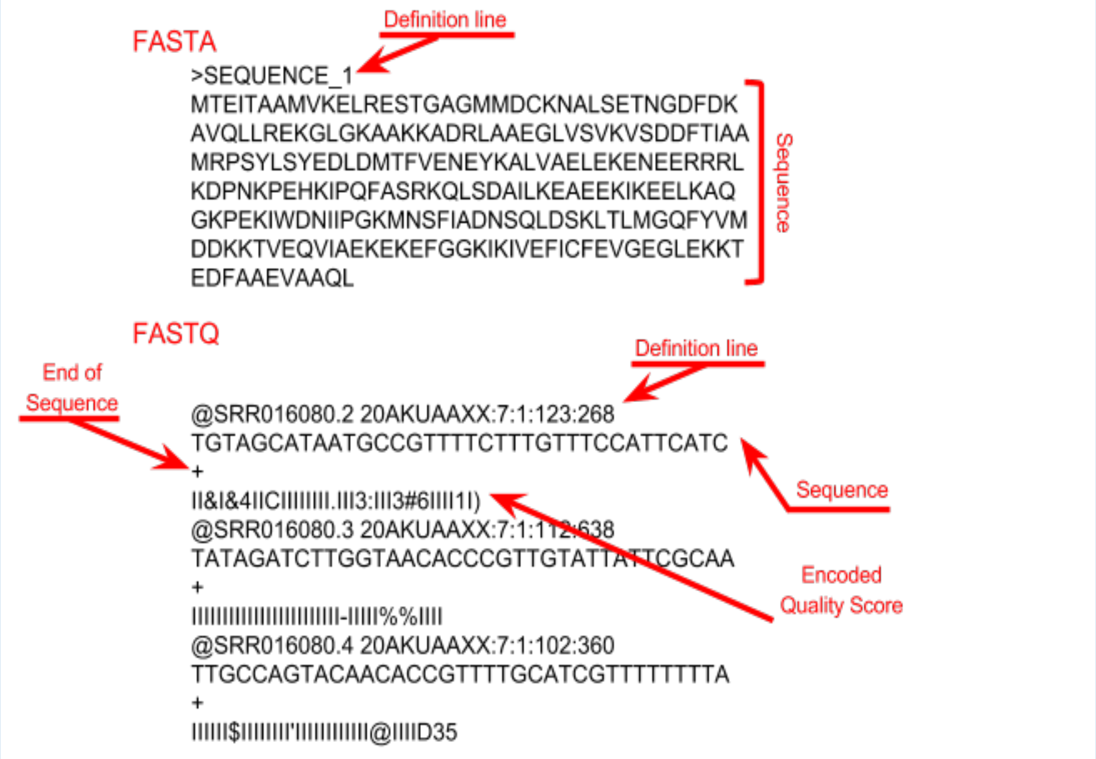
10: Log output of Sickle on data 4 and data 3
9: Singletons from Paired-End output of Sickle on data 4 and data 3
8: Paired-End reverse strand output of Sickle on data 4 and data 3
7: Paired-End forward strand output of Sickle on data 4 and data 3
6: FastQC on data 3: RawData
5: FastQC on data 3: Webpage
4: ftp://ftp.sra.ebi.ac.uk /vol1/fastq/SRR104 /008/SRR1041268/SRR1041268_2.fastq
3: ftp://ftp.sra.ebi.ac.uk /vol1/fastq/SRR104 /008/SRR1041268/SRR1041268_1.fastq
2: ftp://ftp.sra.ebi.ac.uk /vol1/fastq/SRR104 /000/SRR1041270/SRR1041270_2.fastq
1: ftp://ftp.sra.ebi.ac.uk /vol1/fastq/SRR104 /000/SRR1041270/SRR1041270_1.fastq

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



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

```
@SRR016080.2 20AKUAAXX:7:1:123:268
TGTAGCATAATGCCGTTTTCTTTGTTTCCATTCATC
+
II&I&4IICIIIIIIII.III3:III3#6IIII1I)
@SRR016080.3 20AKUAAXX:7:1:112:638
TATAGATCTTGGTAACACCCGTTGTATTATTCGCAA
+
IIIIIIIIIIIIIIIIIIIIIIII-IIIII%%IIII
@SRR016080.4 20AKUAAXX:7:1:102:360
TTGCCAGTACAACACCGTTTTGCATCGTTTTTTTTA
+
IIIIII$IIIIIIII'IIIIIIIIIIII@IIIID35
```