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

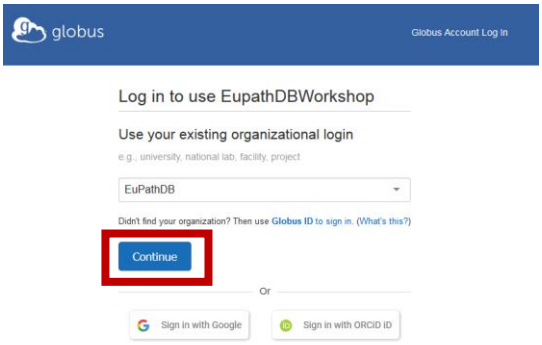
For this exercise, we will retrieve raw sequence files from a repository, assess the quality of the data, and then run the data through a workflow (or pipeline) that will align the data to a reference, calculate expression values and determine differential expression. Part 1, uploading data and starting 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.

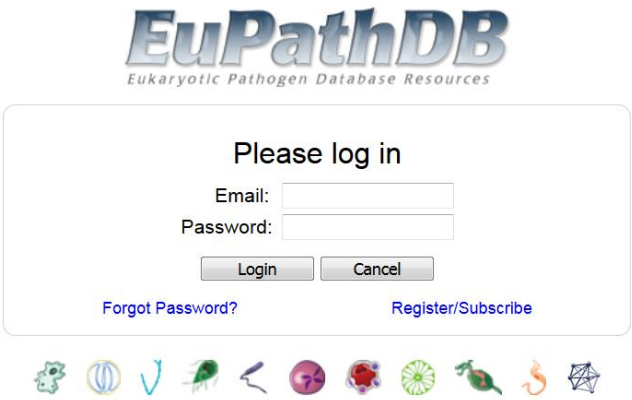
Section I: Setting up your EuPathDB Galaxy account

Step 1: Access the EuPathDB Galaxy instance at the following URL:
http://eupathdbworkshop.globusgenomics.org/

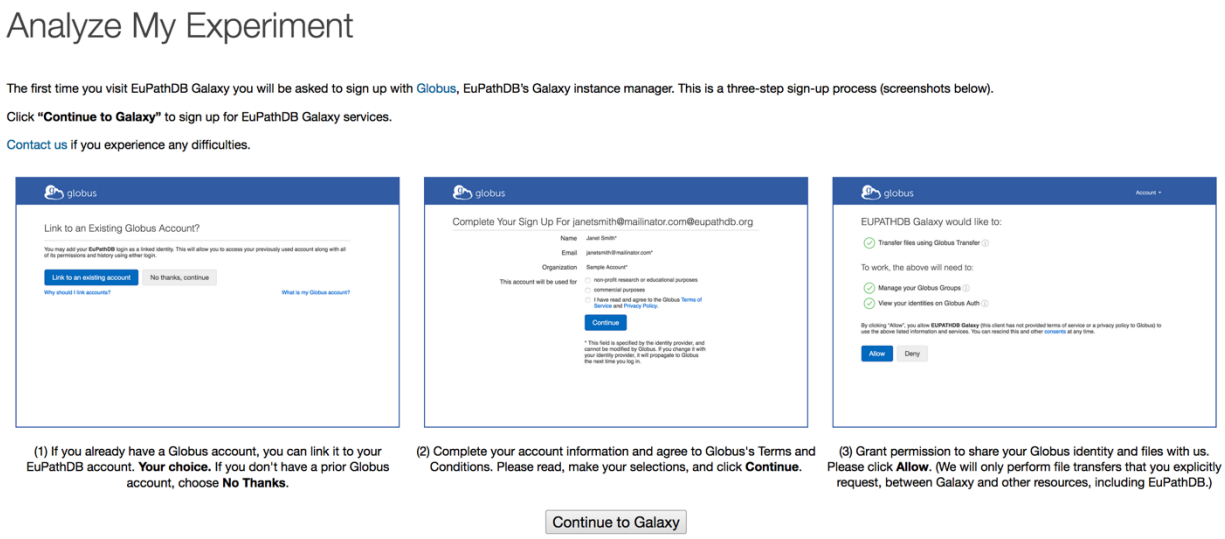
Step 2: On the next page you will be asked to define your organization. Choose EuPathDB and click Continue.



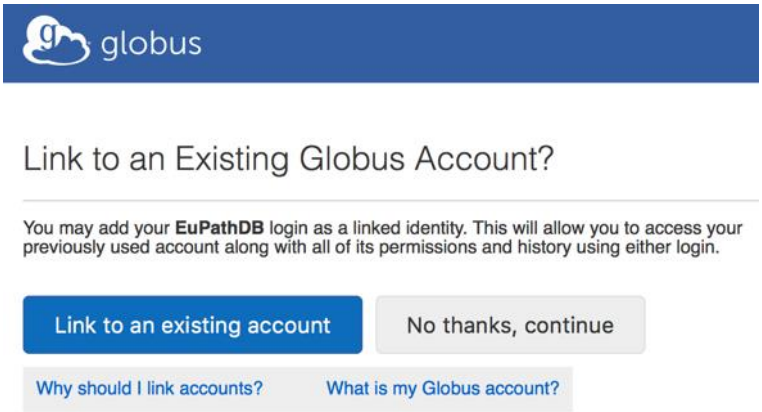
Step 3: Log in to EuPathDB (if you are not logged in already).



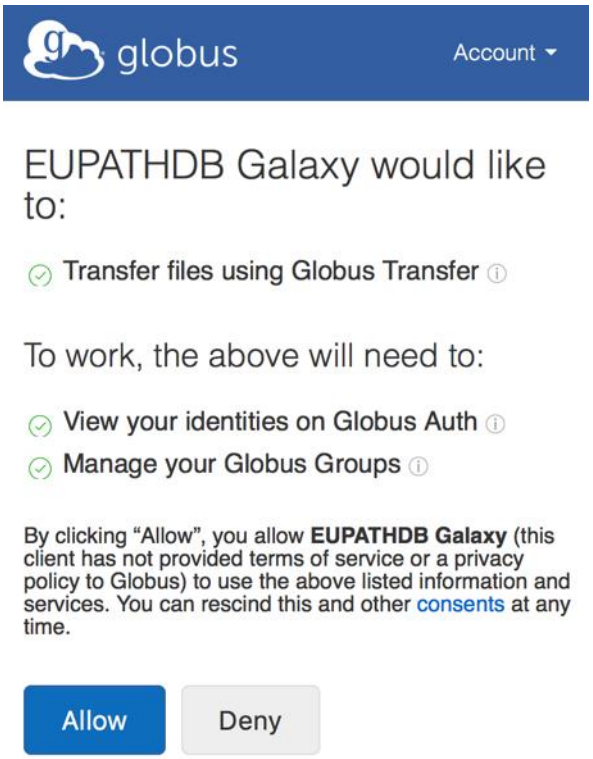
Step 4: Next, sign up for the EuPathDB Galaxy instance.



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



Step 7: Click on “Allow”

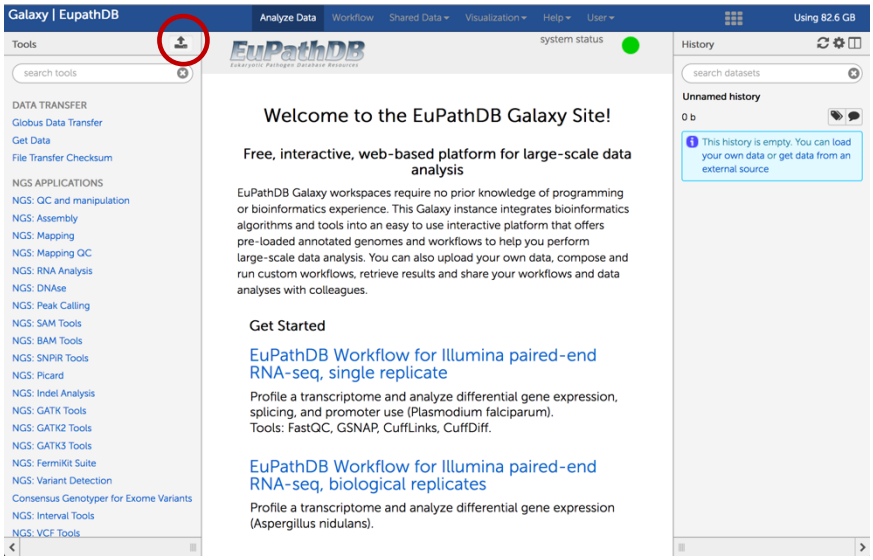


Step 8: Congratulations, you are in!

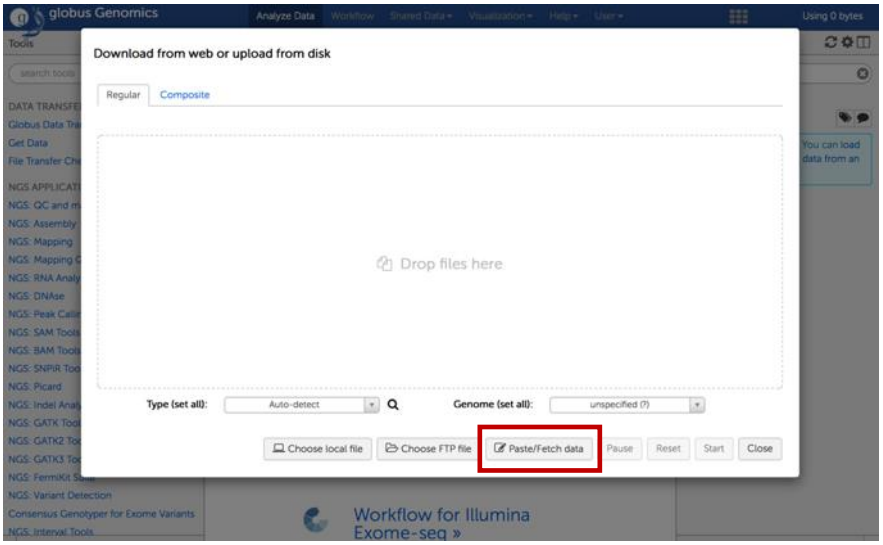
Section II: Importing data to Galaxy

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).

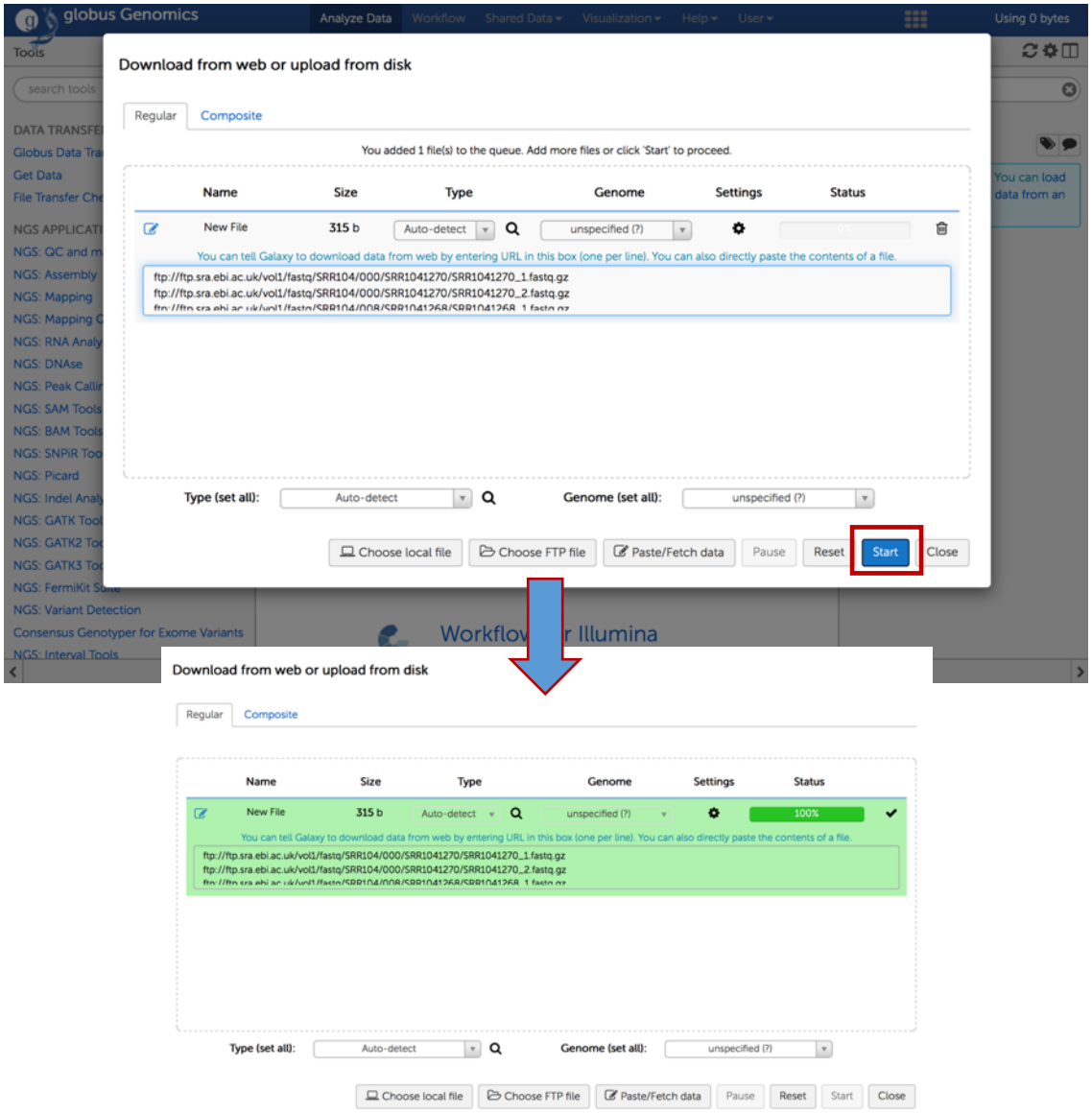
Step 1: Click on the “Get data” icon. This will open up a window that allows you to “**Download from web or upload from disk”**



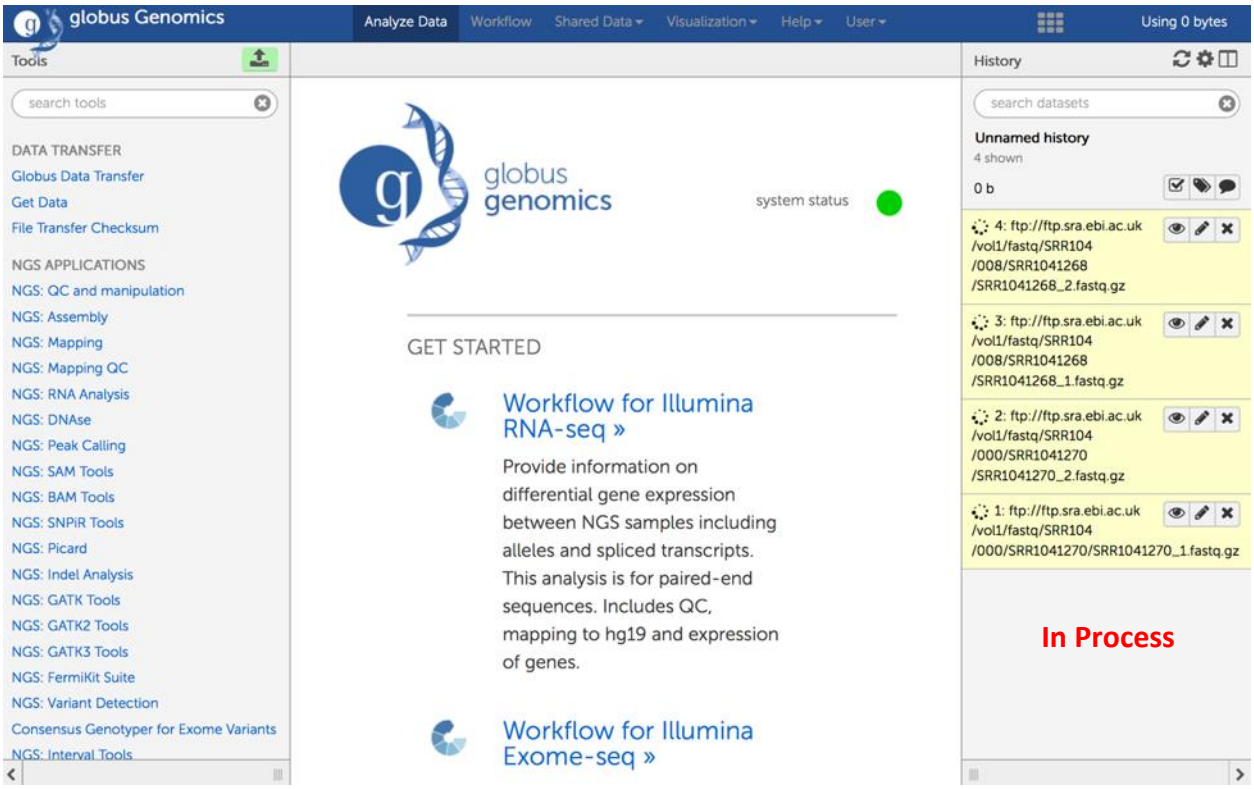
Step 2: In the “Download from web or upload from disk” window click on “Paste/Fetch data”

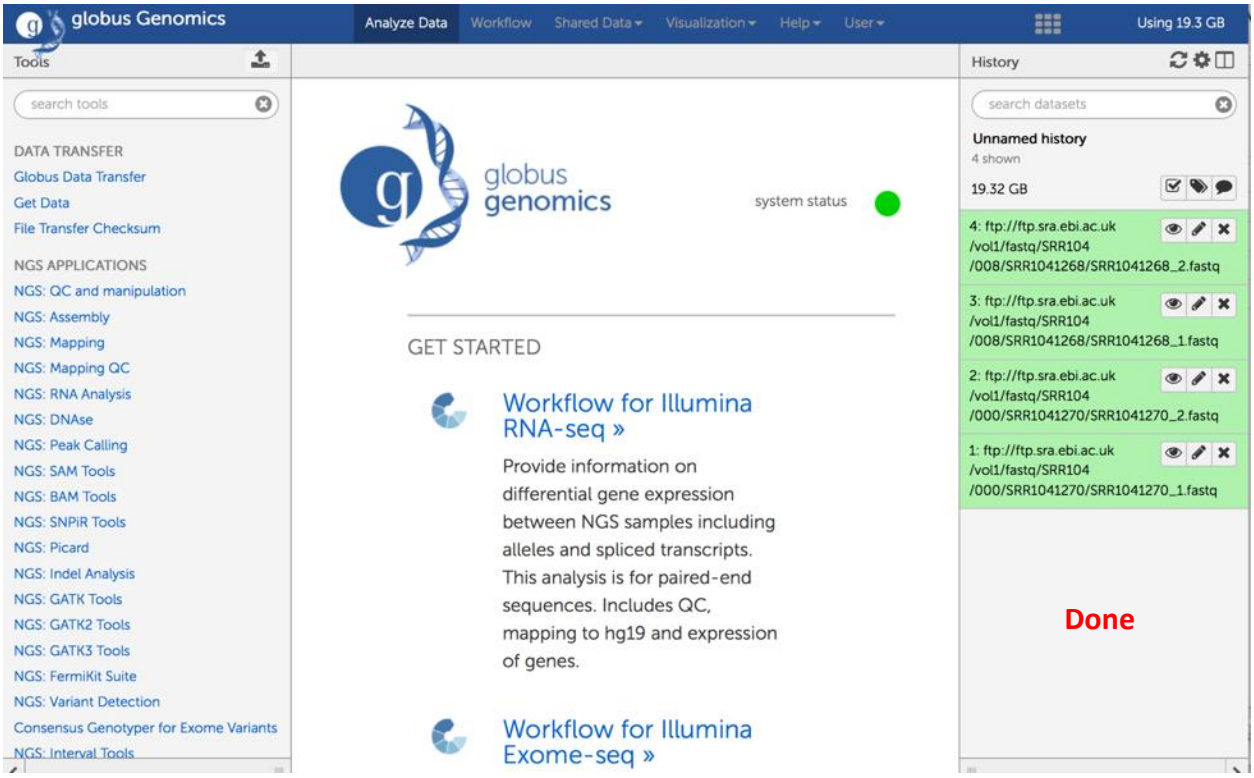


Step 3: Paste the four URLs corresponding to the four files for your group. Links are below in this exercise under Group Assignments. Each URL must be on a new line. Then click on “Start”.



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.


globus Genomics
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 0 bytes
Tools
search tools
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
system status
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 »
History
search datasets
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


globus Genomics
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 19.3 GB
Tools
search tools
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
system status
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 »
History
search datasets
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:

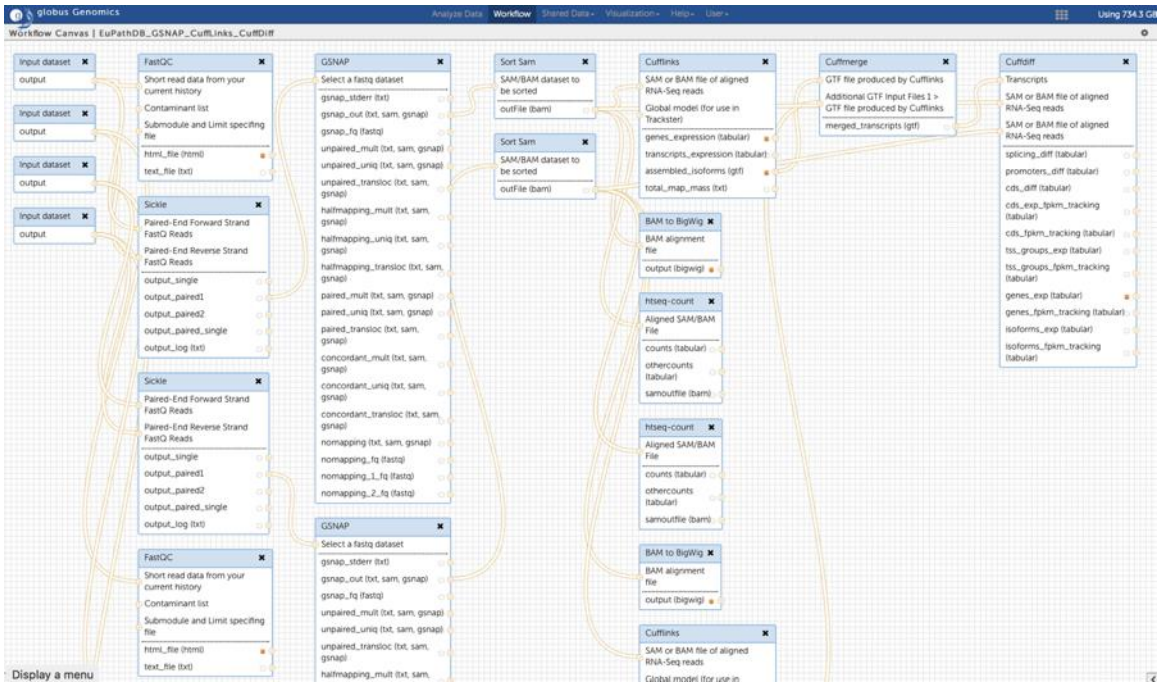
WT:
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR180/001/SRR1805881/SRR1805881_1.fastq.gz
ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR180/001/SRR1805881/SRR1805881_2.fastq.gz

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

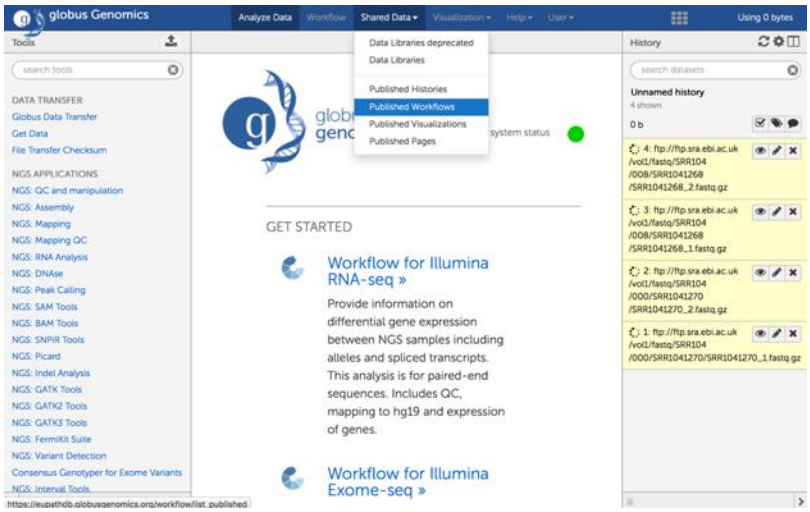
Section II: Running a workflow in Galaxy

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:

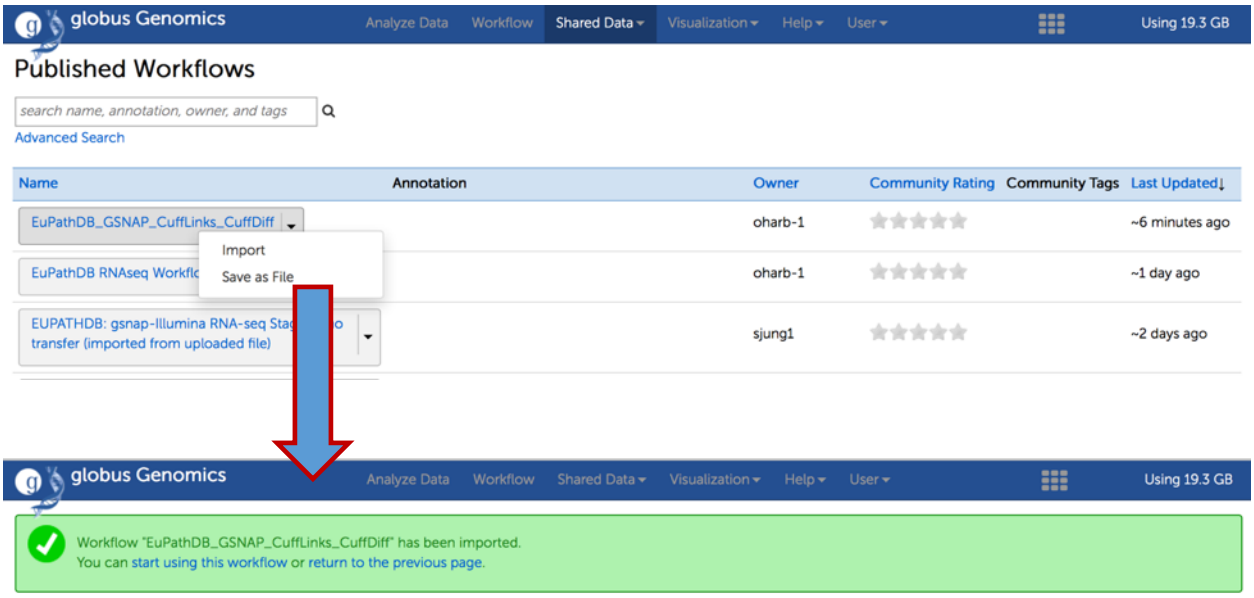
1. Analyzes the reads in your files and generates FASTQC reports.
2. Trims the reads based on their quality scores.
3. Aligns the reads to a reference genome using GSNAP and generates coverage plots.
4. Determines FPKM values for each sample and generates gene/transcript models.
5. Determines differential expression of genes between the samples.



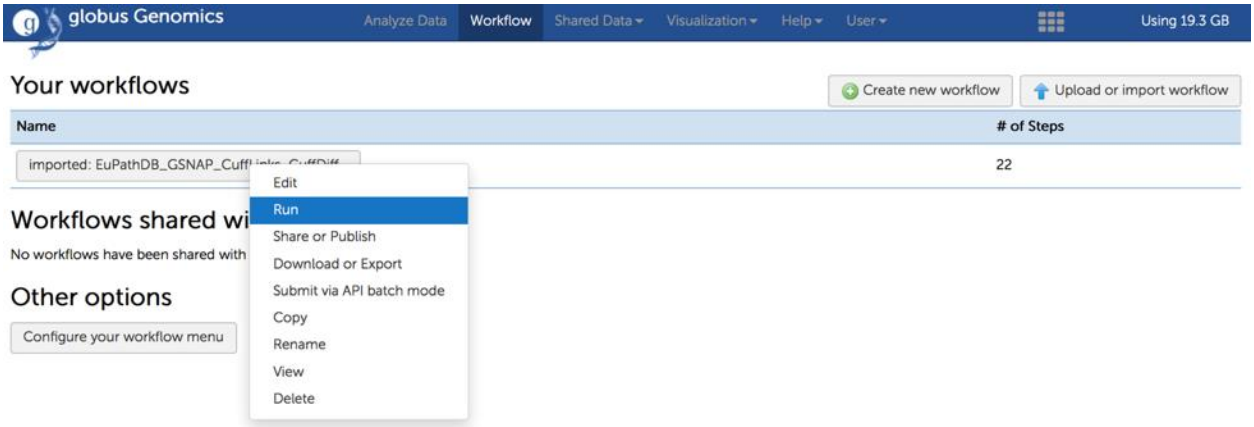
Step 1: Import the workflow called "EuPathDB_GSNAP_CuffLinks_CuffDiff" – click on the shared data menu item and select "Published Workflows" from the menu.



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

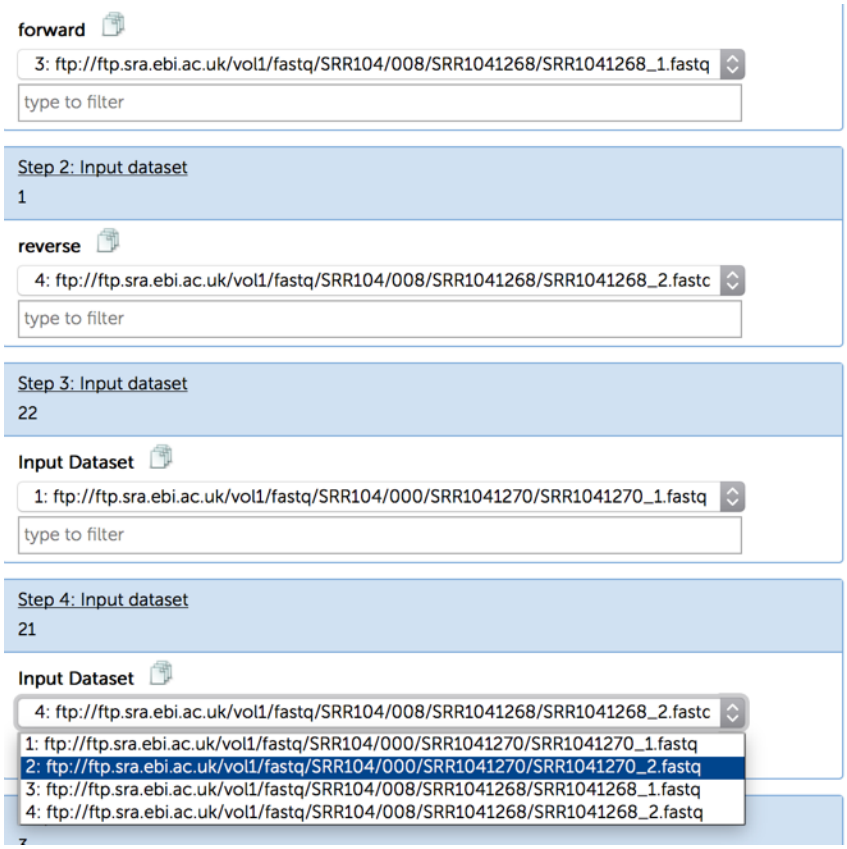


Step 3: Click on "Workflow" in the menu at the top of the page. On the next page click on the arrow next to your imported workflow and select the "Run" option.

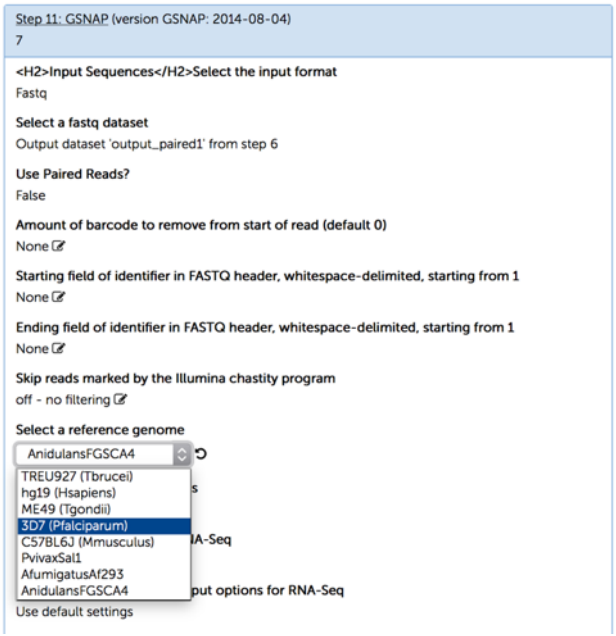


Step 4: Configure your workflow – there are multiple steps in the workflow but you do not need to configure all of them. For the purpose of this exercise you will need to configure the following:

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.

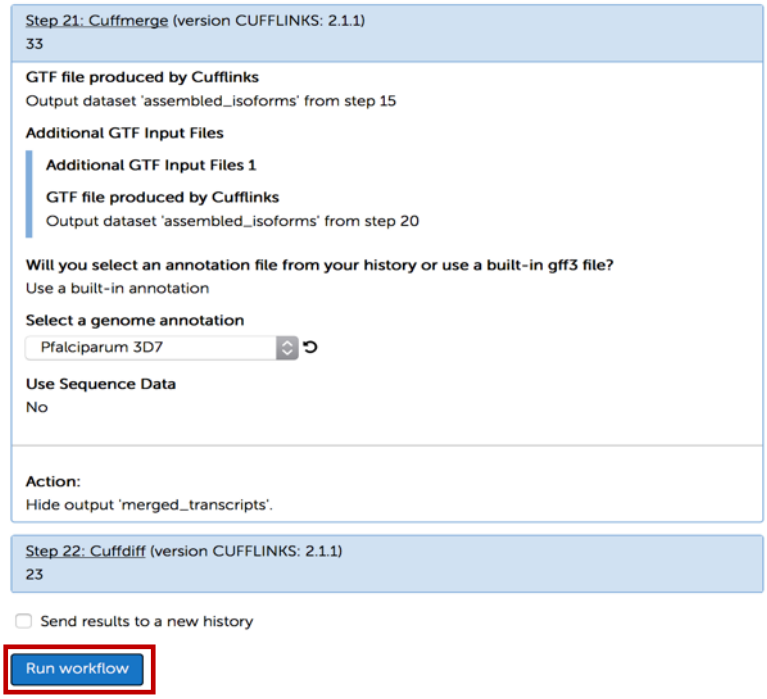


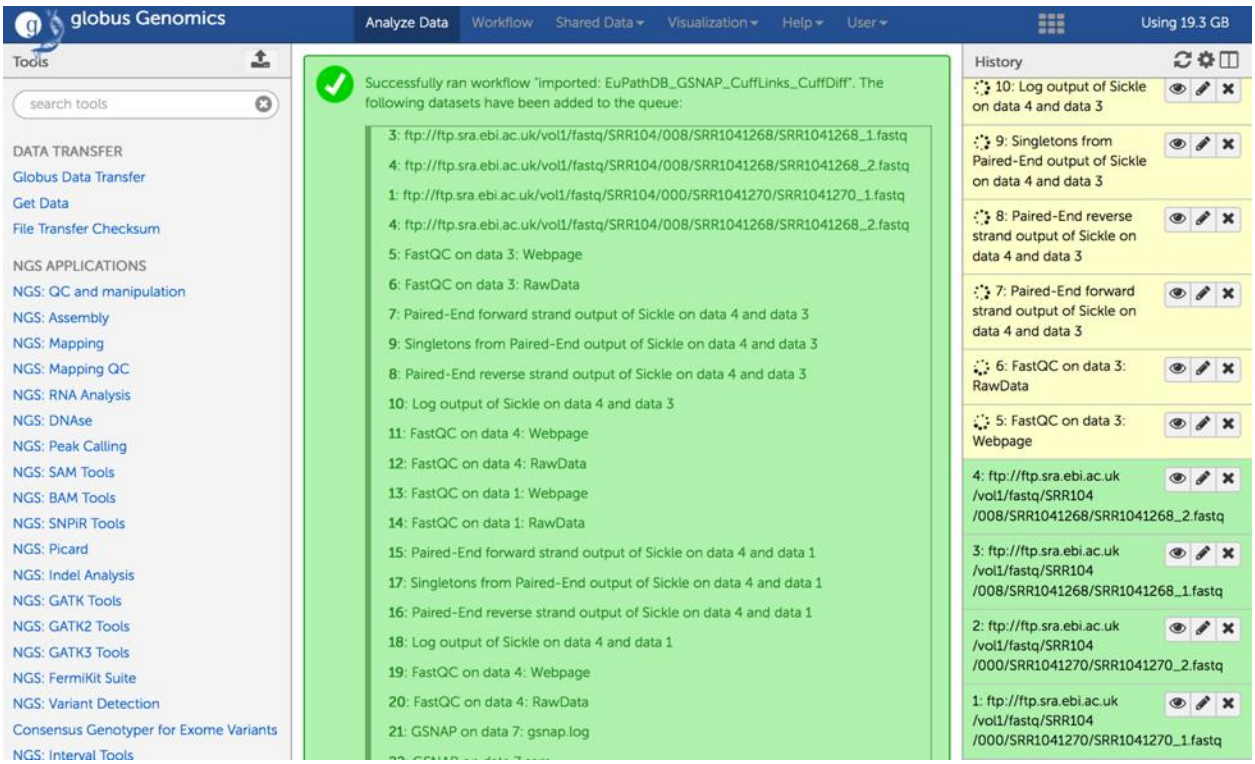
b. Scroll down to steps 11 and 12 (GSNAP). Click on the name of the step to open up the parameters. Select the correct reference organism in each of the steps.



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"





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