Variant Calling analysis, Part I: Uploading data and starting the workflow (Group Exercise)

Learning objectives:

- Import data from EBI to the VEuPathDB Galaxy
- Run a pre-configured SNP workflow

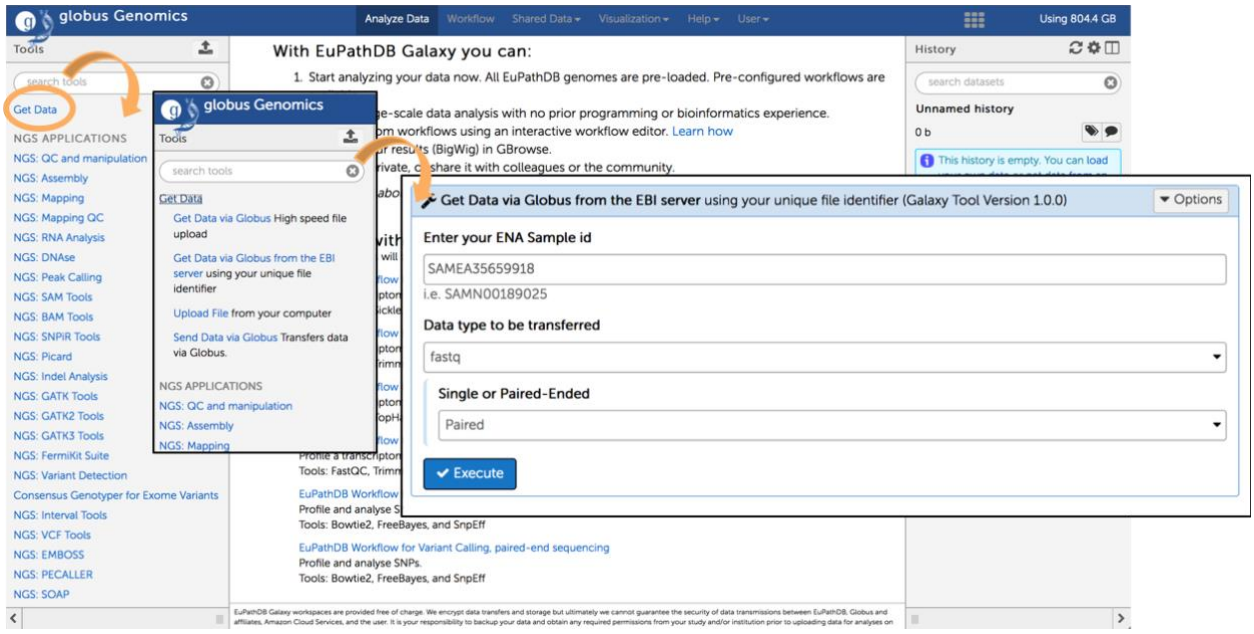
In this exercise we will work in groups to retrieve DNA sequence data from the sequence repository and analyze it for variants using a workflow in VEuPathDB Galaxy.

To upload the data, we will use the same approach as in RNA-Seq Galaxy analysis section. 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.

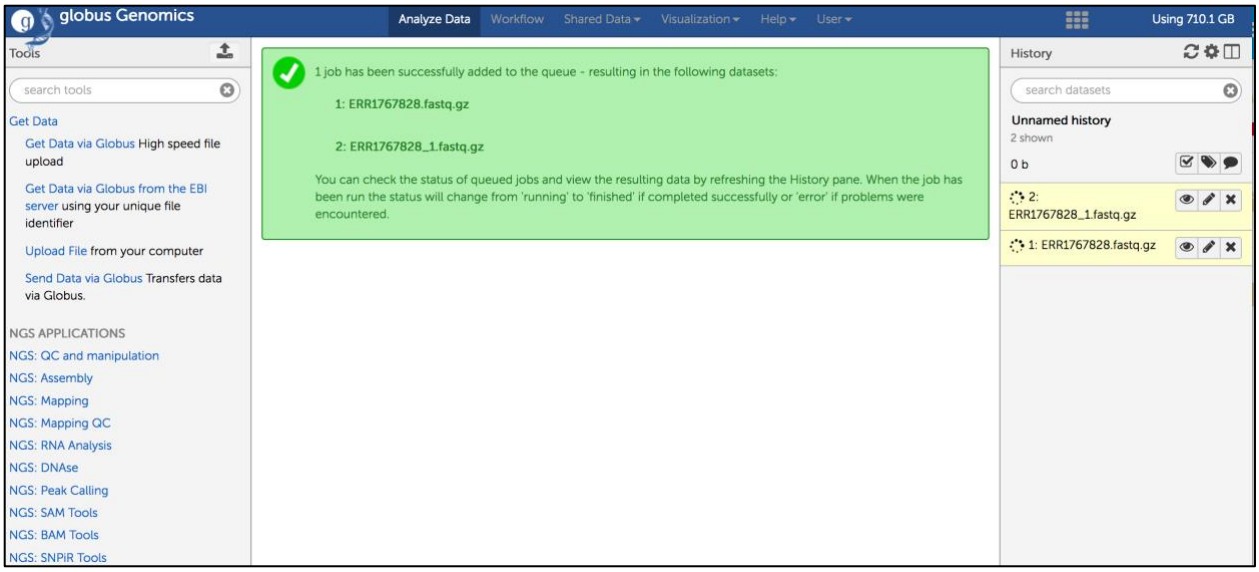
1. Click on the *Get Data* link.
2. Click on the *Get Data via Globus from the EBI server* link.
3. The next window allows you to enter the sample ID. This ID starts with the letters 'SAM'. Choose the sample ID for your group from the list below and use it in this form.

Note: it is very important that you select whether the data is single or paired-ended.

4. Once the form is properly filled, click on the *Execute* button to start the data transfer process.



If you click on the "Upload File from your computer" you will be able to use url links to initiate file download.



Group assignments:

Groups 1 & 2 will be examining *Aspergillus fumigatus* isolates F18085 & AF90-PE-ILI27-01. Paired-end sequencing.
The data is available in the sequence repositories:
https://www.ebi.ac.uk/ena/browser/view/PRJNA287921

Group number	1	2
Sample Name	F18085	AF90-PE-ILI27-01
Sample Accession Numbers	SAMN01815907	SAMN01815887
Ref genome in Galaxy	FungiDB-29_AfumigatusAf293	

Group 3 & 4 will be examining Australian pathotypes of *Puccinia graminis* f. sp *tritici* PGT326 & PGT21-0. Paired-end sequencing.
https://www.ebi.ac.uk/ena/browser/view/PRJNA415866

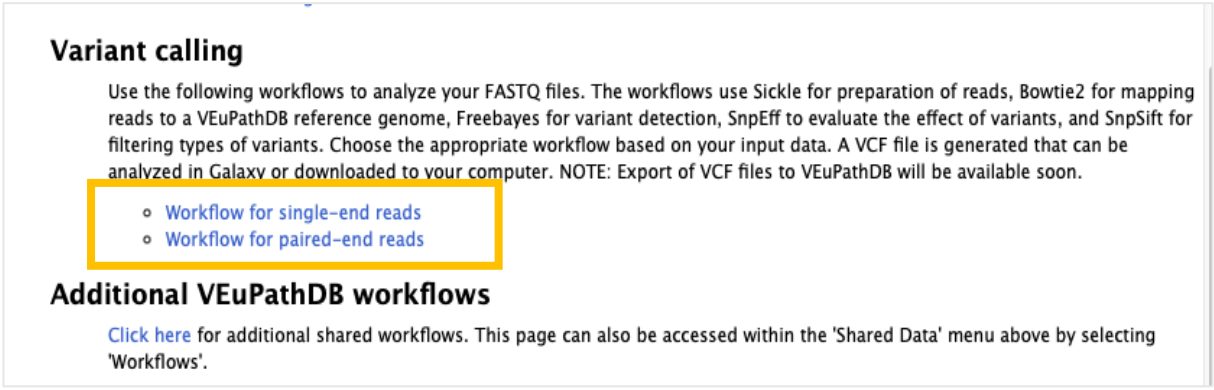
Group number	3	4
Sample Name	PGT326	PGT21-0
Sample Accession Numbers	SAMN07836890	SAMN07836889
Ref genome in Galaxy	FungiDB-29_PgraminisCLR75-36-700-3	

Group 5 & 6 will be examining *Candida auris* isolates from Australia and California, respectively. Paired-end sequencing.
NOTE: GROUP 6 will have to transfer data using the URL (Click Get data > Upload file > Paste/Fetch data. Copy and paste 2 URLs (see below)).
https://www.ebi.ac.uk/ena/browser/view/PRJNA485022 and
https://www.ebi.ac.uk/ena/browser/view/PRJNA480539

Group number	5	6
Sample Name	Au	USA_Cali
Sample Accession Numbers	SAMN09780833	ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR750/009/SRR7507279/SRR7507279_1.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR750/009/SRR7507279/SRR7507279_2.fastq.gz
Ref genome in Galaxy	FungiDB-37_CaurisB8441	

Running a variant calling workflow:

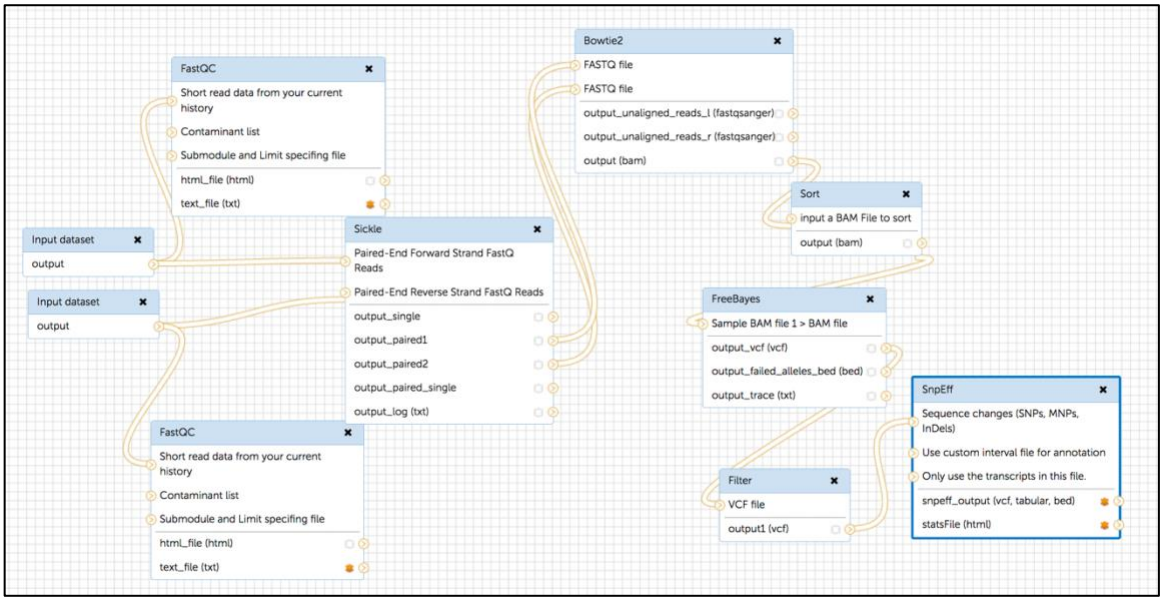
- Once the data files have been transferred into your galaxy history you need to choose an appropriate workflow. EuPathDB provides some preconfigured workflows on the EuPathDB Galaxy instance home page.
- Remember to choose the appropriate workflow –for paired-end samples.



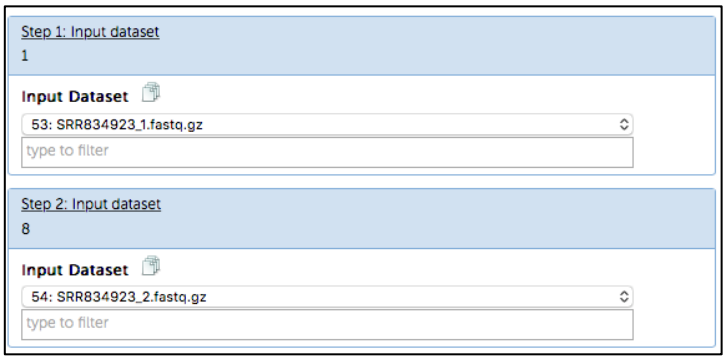
The pre-configured workflows follow these steps:

- Determine quality of the reads in your files and generates FASTQC reports
- Trim reads based on their quality scores
- Align reads to a reference genome using Bowtie2 and generating coverage plots
- Sort alignments with respect to their chromosomal positions
- Detect variants using FreeBayes
- Filter SNP candidates

- Analyze and annotate of variants, and calculation of the effects via SnpEff



- Next, set workflow parameters.
  - Make sure that the input steps for paired-end are set to the *xxxx_1.fastq.gz* and *xxxx_2.fastq.gz* as by default both have the same one selected.



- Select the correct reference genome (for steps: Bowtie2, FreeBayes, SnpEff)
- Click on the *Run Workflow* button.