Variant calling in VEuPathDB galaxy (Part 1)

Learning objectives:

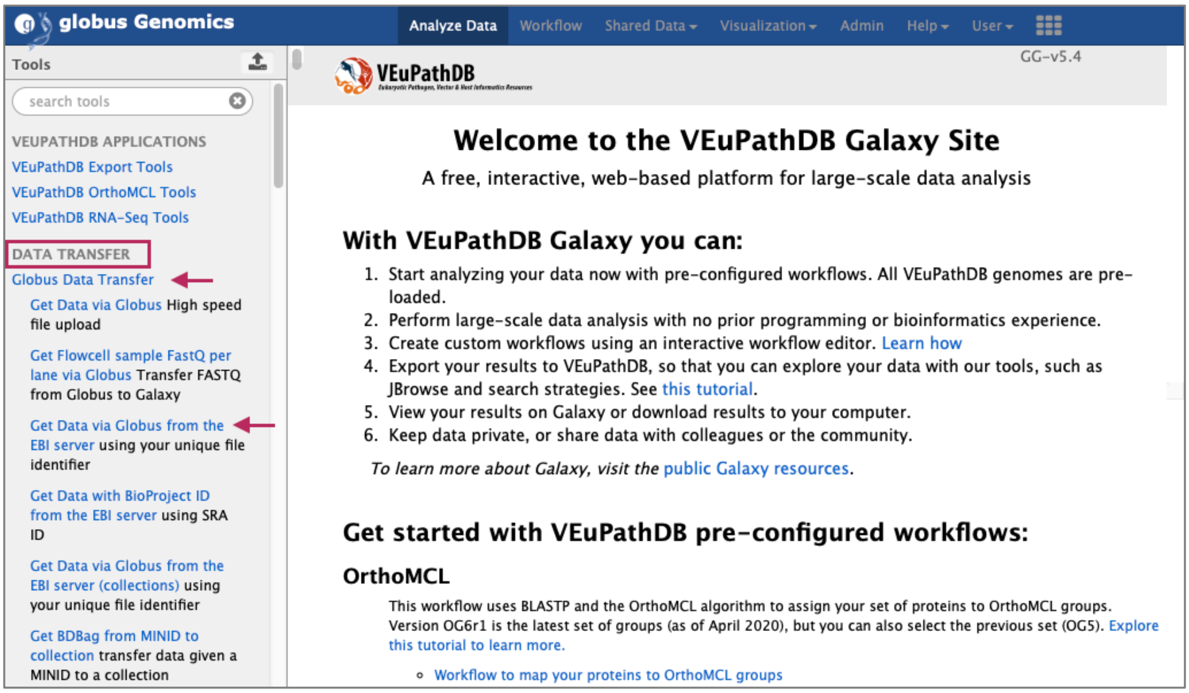
1. Retrieve DNA sequence data from the sequence repository EBI and upload data to VEuPathDB Galaxy using Globus Data Transfer;
2. Name a new project/history;
3. Deploy a Variant calling workflow in the VEuPathDB Galaxy.

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. VEuPathDB developed its Galaxy instance in collaboration with Globus Genomics (VEuPathDB Galaxy). To learn how to use Galaxy, follow this link to access tutorials prepared by the Galaxy Training Network: https://wiki.galaxyproject.org/Learn#Galaxy_101

There are different ways to get data into Galaxy. In this exercise we will use Globus Data Transfer to get data from the EBI server using a unique project ID.

1. Retrieve DNA sequence data from the sequence repository and upload data to VEuPathDB Galaxy using Globus Data Transfer option.

a. Click on the “Globus Data Transfer” menu on the left to expand the Data Transfer section.
b. Click on the “Get Data via Globus from the EBI server” link.



VEUPATHDB APPLICATIONS
VEuPathDB Export Tools
VEuPathDB OrthoMCL Tools
VEuPathDB RNA-Seq 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

Welcome to the VEuPathDB Galaxy Site

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

With VEuPathDB Galaxy you can:

1. Start analyzing your data now with pre-configured workflows. All VEuPathDB genomes are pre-loaded.
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. Export your results to VEuPathDB, so that you can explore your data with our tools, such as JBrowse and search strategies. See this tutorial.
5. View your results on Galaxy or download results to your computer.
6. Keep data private, or share data with colleagues or the community.

To learn more about Galaxy, visit the public Galaxy resources.

Get started with VEuPathDB pre-configured workflows:

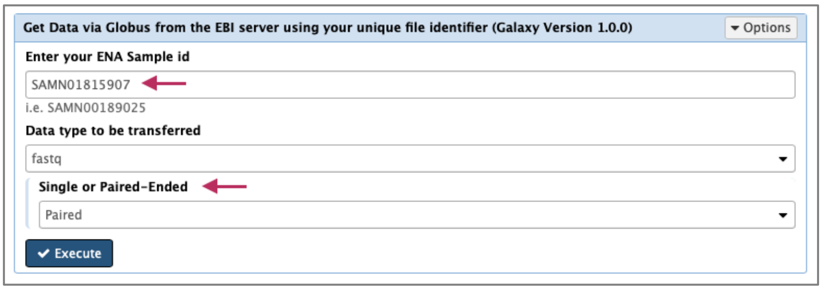
OrthoMCL

This workflow uses BLASTP and the OrthoMCL algorithm to assign your set of proteins to OrthoMCL groups. Version OG6r1 is the latest set of groups (as of April 2020), but you can also select the previous set (OG5). Explore this tutorial to learn more.

- Workflow to map your proteins to OrthoMCL groups

c. Enter ENA sample ID and define the dataset type to be transferred into the VEuPathDB Galaxy workspace.

The ENA ID should start with the letters ‘SAM’. For this exercise, we will use SAMN01815907, which is a paired-ended dataset. Take care to specify whether a dataset is a single or paired-ended as incorrect selection will cause the upload to fail.

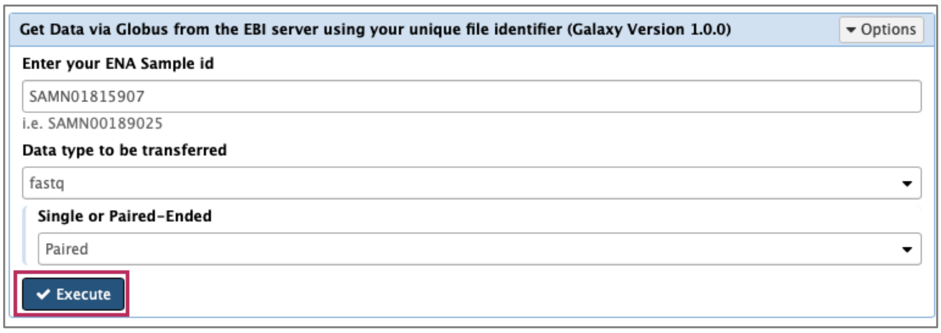

Get Data via Globus from the EBI server using your unique file identifier (Galaxy Version 1.0.0)

Enter your ENA Sample id: SAMN01815907
i.e. SAMN00189025

Data type to be transferred: fastq

Single or Paired-Ended: Paired

d. Once the form is properly filled, click on the “Execute” button to start the data transfer process.


Get Data via Globus from the EBI server using your unique file identifier (Galaxy Version 1.0.0)

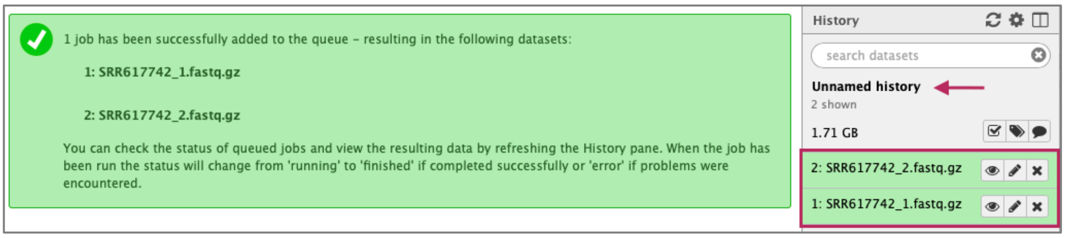
Enter your ENA Sample id: SAMN01815907
i.e. SAMN00189025

Data type to be transferred: fastq

Single or Paired-Ended: Paired

e. When the job has been successfully deployed and added to the queue, the screen will refresh, and the added job will appear in the history on the right.

Note: new jobs are highlighted in grey, in progress – yellow, and those completed are in green.


1 job has been successfully added to the queue – resulting in the following datasets:

1: SRR617742_1.fastq.gz

2: SRR617742_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
1.71 GB

2: SRR617742_2.fastq.gz

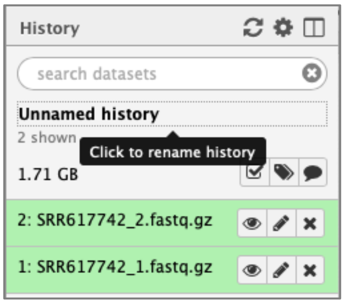
1: SRR617742_1.fastq.gz

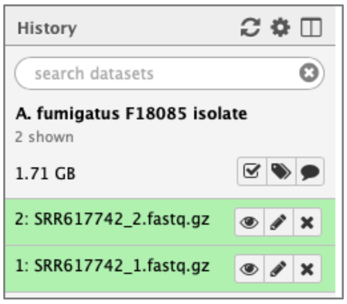
Notice that there are two files appearing in the history on the right. This is because the uploaded data is paired-ended.

2. Rename your history.

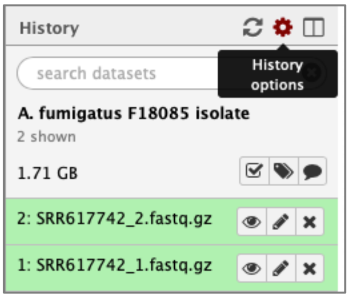
By default, all new jobs will be added to the current history on the right. Unless renamed, the history will show up in your history as “Unnamed history”. Let’s rename the history to help us track this project in the future.

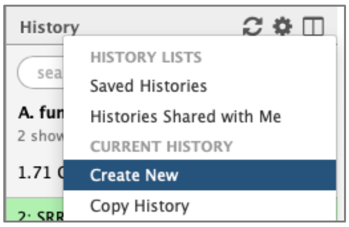
a. Click on the “Unnamed history” and type “A. fumigatus F18085 isolate”, and then press “enter” to rename this history.





Note: if you would like to start a new project/history, click on the wheel button at the top of the history section and select “Create a new history”.

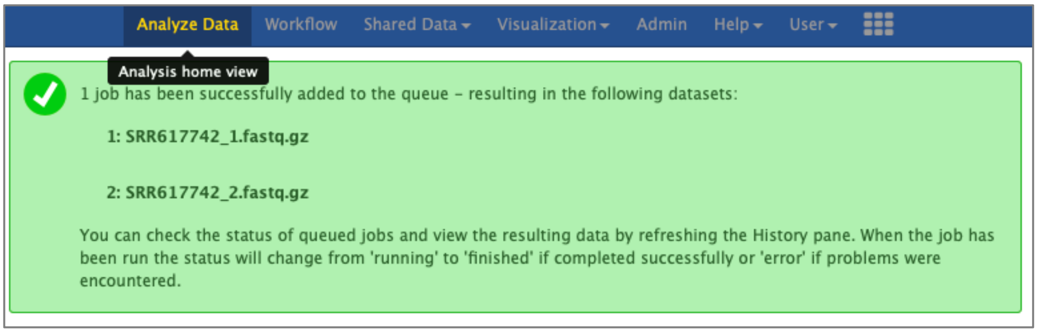




3. Deploy a Variant calling workflow.

VEuPathDB Galaxy main landing page has several workflows for variant calling.

a. To navigate to the main page, click on the “Analyze Data”, which is located in the main menu at the top.



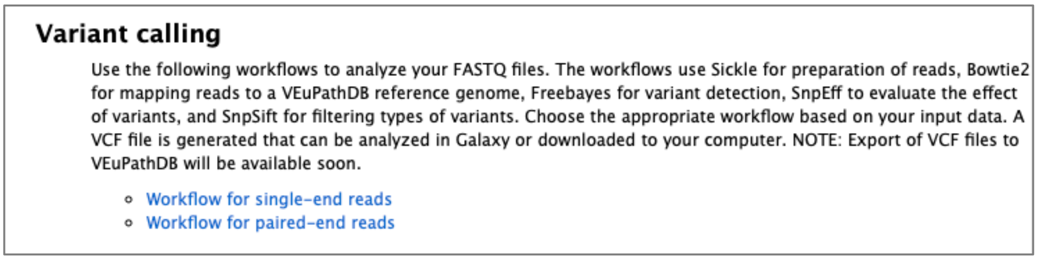
1 job has been successfully added to the queue – resulting in the following datasets:

1: SRR617742_1.fastq.gz

2: SRR617742_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.

b. Scroll down to the Variant calling section and choose the workflow for paired-end reads.



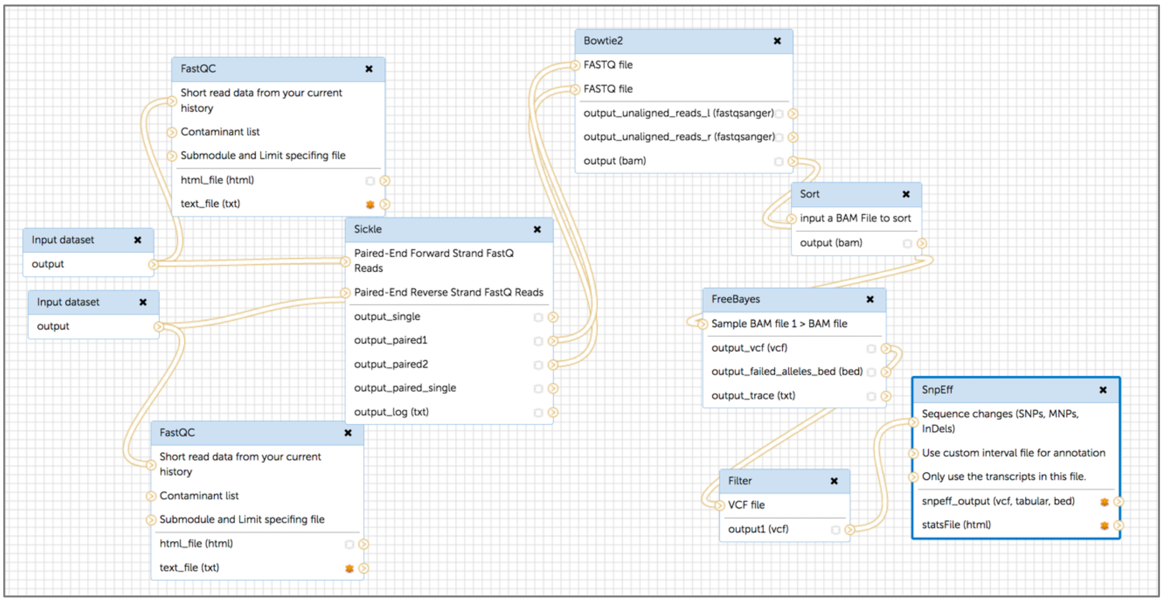
Variant calling

Use the following workflows to analyze your FASTQ files. The workflows use Sickle for preparation of reads, Bowtie2 for mapping reads to a VEuPathDB reference genome, Freebayes for variant detection, SnpEff to evaluate the effect of variants, and SnpSift for filtering types of variants. Choose the appropriate workflow based on your input data. A VCF file is generated that can be analyzed in Galaxy or downloaded to your computer. NOTE: Export of VCF files to VEuPathDB will be available soon.

- Workflow for single-end reads
- Workflow for paired-end reads

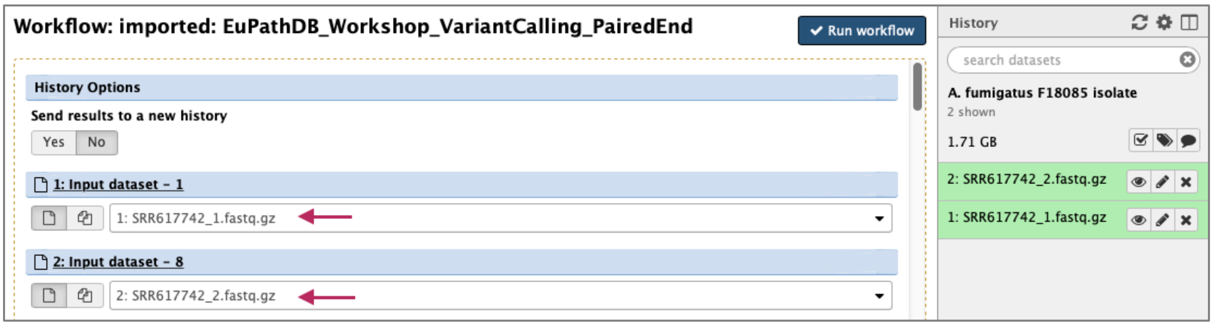
The pre-configured variant calling workflows include the following steps:

- Determine quality of the reads and generate reports (FastQC);
- Trim reads based on their quality scores (Sickle);
- Align reads to a reference genome using Bowtie2 and generate coverage plots ;
- Sort alignments with respect to their chromosomal positions (Sort);
- Detect variants (FreeBayes);
- Filter SNP candidates (Filter);
- Analyze and annotate variants, and calculate the effects of SNPs via SnpEff.


FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
Input dataset
output
Input dataset
output
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)
FastQC
Short read data from your current history
Contaminant list
Submodule and Limit specifing file
html_file (html)
text_file (txt)
Bowtie2
FASTQ file
FASTQ file
output_unaligned_reads_l (fastqsanger)
output_unaligned_reads_r (fastqsanger)
output (bam)
Sort
input a BAM File to sort
output (bam)
FreeBayes
Sample BAM file 1 > BAM file
output_vcf (vcf)
output_failed_alleles_bed (bed)
output_trace (txt)
Filter
VCF file
output1 (vcf)
SnpEff
Sequence changes (SNPs, MNPs, InDels)
Use custom interval file for annotation
Only use the transcripts in this file.
snpeff_output (vcf, tabular, bed)
statsFile (html)

c. Click on the workflow for paired-end reads and set workflow parameters.

- Make sure that the input steps for paired-end data are set to the xxxx_1.fastq.gz and xxxx_2.fastq.gz file (by default the same file will be selected in both fileds).



Workflow: imported: EuPathDB_Workshop_VariantCalling_PairedEnd

✓ Run workflow

History Options

Send results to a new history

Yes No

1: Input dataset - 1

1: SRR617742_1.fastq.gz

2: Input dataset - 8

2: SRR617742_2.fastq.gz

History

search datasets

A. fumigatus F18085 isolate

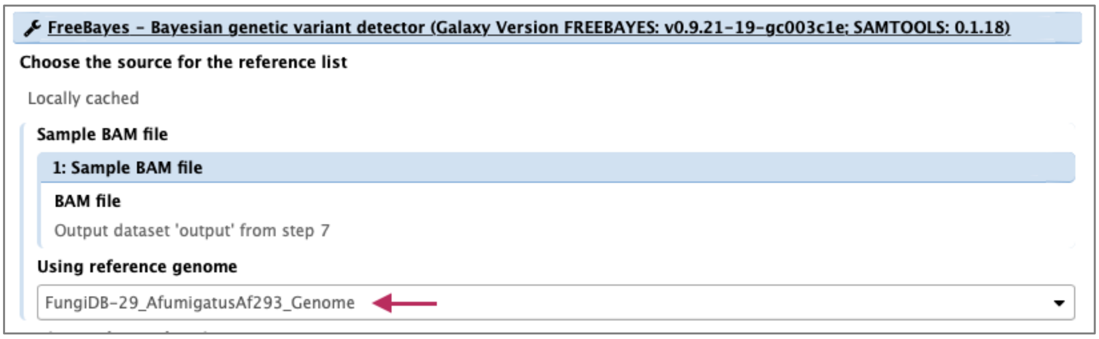
2 shown

1.71 GB

2: SRR617742_2.fastq.gz

1: SRR617742_1.fastq.gz

- Select the correct reference genome.
  - Select *Aspergillus fumigatus* Af293 as a reference genome (steps: Bowtie2, FreeBayes, SnpEff).



FreeBayes - Bayesian genetic variant detector (Galaxy Version FREEBAYES: v0.9.21-19-gc003c1e; SAMTOOLS: 0.1.18)

Choose the source for the reference list

Locally cached

Sample BAM file

1: Sample BAM file

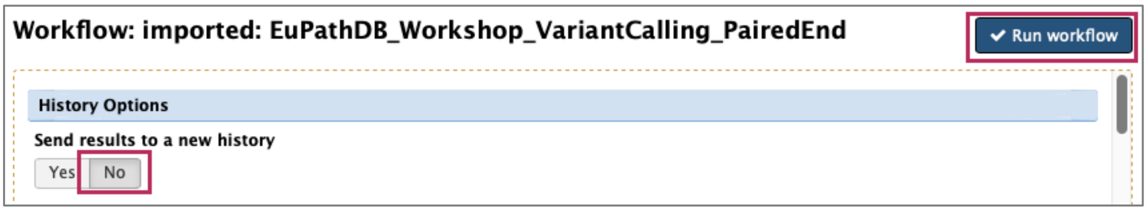
BAM file

Output dataset 'output' from step 7

Using reference genome

FungiDB-29_AfumigatusAf293_Genome

- Choose to deploy the analysis within the same history and click on the Run workflow button.



Workflow: imported: EuPathDB_Workshop_VariantCalling_PairedEnd

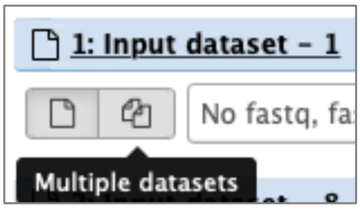
✓ Run workflow

History Options

Send results to a new history

Yes [✓ No]

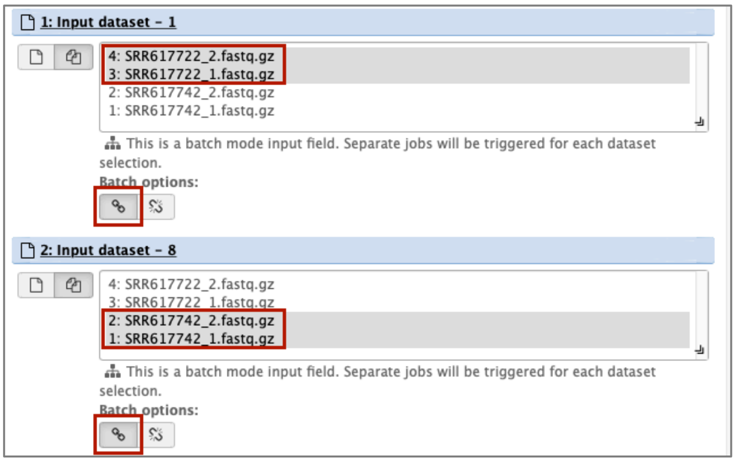
Note: You can use the same workflow to analyze multiple samples in batches. The. Upload steps remain the same, however, when setting up the workflow, click on multiple dataset button within the input dataset section.



1: Input dataset - 1

No fastq, fa

Multiple datasets



1: Input dataset - 1

4: SRR617722_2.fastq.gz
3: SRR617722_1.fastq.gz
2: SRR617742_2.fastq.gz
1: SRR617742_1.fastq.gz

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Batch options:

2: Input dataset - 8

4: SRR617722_2.fastq.gz
3: SRR617722_1.fastq.gz
2: SRR617742_2.fastq.gz
1: SRR617742_1.fastq.gz

This is a batch mode input field. Separate jobs will be triggered for each dataset selection.

Batch options: