

easyMF User Mannual

(version 1.0)

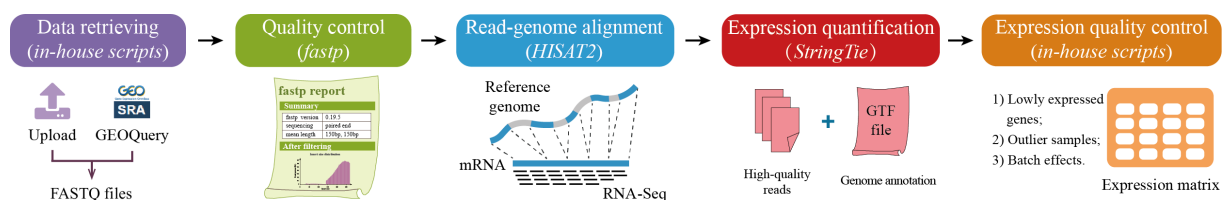
easyMF is a user-friendly web platform that aims to facilitate biological discovery from large-scale transcriptome data through matrix factorization (MF). It offers several functional tools for gene expression matrix generation, expression matrix factorization, and metagene-based exploratory analysis including sample clustering, signature gene identification, and functional gene discovery. It should be noted that the metagenes in PCA can be ranked by the extent to which they explain the variation in the data, the features in both ICA and NMF are assumed to have equal weight. Therefore, we noticed the user that the choice of MF techniques should be based on the purpose of one study with a preferred choice: PCA finds sources of separation in the data, which resulting in identification of dominated variation; ICA learns variation that are statically independent, which resulting in more accurate literature-derived association; NMF is well suited to transcriptional data, which is typically non-negative itself.

- easyMF project is hosted on <https://github.com/cma2015/easyMF>.
- easyMF docker image is available in <https://hub.docker.com/r/malab/easymf>.
- easyMF demo server can be accessed via <http://easymf.omicstudio.cloud>.
- The following part shows installation of easyMF docker image and detailed documentation for each function in easyMF.

0. Matrix Preparation

Matrix factorization is typically started with an input of a gene expression matrix (genes in rows and individual samples in columns), which prompt us to design this module including three functions to prepare a high-quality gene expression matrix for downstream analysis.

The gene expression matrix can be automatically generated from raw reads using a bioinformatics pipeline (see following figure).



This module consists of three functions: **Data Retrieval**, **Expression Matrix Generation** and **Expression Quality Control**.

Functions/Tools	Description	Inputs	Outputs	Time (test data)	Program	References
Data Retrieval	Retrieve genome sequences, genome annotation, and RNA-Seq data automatically from public databases	Select a species; Database version; Data type	Genome sequences (in terms of Reference genome sequence); Genome annotation (in terms of Reference genome annotation); RNA-Seq data (in terms of Raw RNA-Seq data)	Depends on network speed	In-house scripts	This study
Expression Matrix Generation	Generate a gene expression matrix (genes in rows and individual samples in columns) through raw RNA-Seq quality control, read-genome alignment, and gene expression abundance calculation	Genome sequence and annotation; RNA-Seq data	Gene expression matrix	~ 2 mins	fastp (Raw RNA-Seq quality control)	Chen et al., 2018
					HISAT2 (Read-genome alignment)	Kim et al., 2015
					StringTie (Gene expression abundance calculation)	Pertea et al., 2015
					In-house scripts	This study
Expression Quality Control	Generate a high-quality gene expression matrix through removing lowly expressed genes, outlier samples, or batch effects	Raw gene expression matrix	High-quality gene expression matrix	~ 10s	In-house scripts (Removing lowly expressed genes and outlier samples)	This study

Functions/Tools	Description	Inputs	Outputs	Time (test data)	Program	References
					sva (Removing batch effects)	Leek et al., 2012

1. Data Retrieval

Data Retrieval can be used to retrieve **Genome sequences** and **Genome annotation** from [Ensembl Plants](#), **RNA-Seq data** from [NCBI](#) (National Center for Biotechnology Information) GEO (Gene Expression Omnibus) or SRA (Short Read Archive) databases.

Inputs

For retrieving **genome sequences and annotation**, users need to select option **Obtain Genome Sequences and Annotation**.

- **Select a species:** This option provides the Latin name of 61 species.
- **Database version:** Ensembl releases from 25 to 47 are listed.
- **Data type:** Genome sequences (.fasta) or annotation (.gtf).

For retrieving **RNA-Seq data**, users need to select option **Obtain RNA-Seq data**.

- **Fetch data through data ID or ftp address:** easyMF provides two ways to download RNA-Seq data.
If users select **Fetch data through data ID**, easyMF downloads RNA-Seq data by NCBI's tool *sra-toolkit* (version 2.3.5) through RNA-Seq IDs (such as SRR1765337).
If users select **Fetch data through data address**, easyMF downloads RNA-Seq data by wget using HTTP/FTP addresses.

Outputs

For **Obtain Genome Sequences and Annotation**

- **Reference genome sequence**
- **Reference genome annotation**

For **Obtain RNA-Seq data**

- **Raw RNA-Seq data**

How to use this function

- The following screenshot shows us how to download genome sequences and annotation using easyMF.

Galaxy Analyze Data Workflow Shared Data Visualization Help Login or Register Using 2.0 GB

Tools search tools

Get Data Upload File from your computer

easyMF

- I) MATRIX PREPARATION
- Data Retrieval
- Expression Matrix Generation
- Expression Quality Control
- II) MATRIX FACTORIZATION
- Matrix Factorization
- III) METAGENE-BASED DEEP MINING USING AM
- Functional Gene Discovery
- IV) METAGENE-BASED DEEP MINING USING PM
- Sample Clustering Analysis
- Temporal-spatial Analysis

Workflows

- All workflows

Data Retrieval (Galaxy Version 17.09)

Genome or RNA-Seq → **Step 1: Select "Obtain Genome Sequences and Annotation"**

Obtain Genome Sequences and Annotation

Select a species → **Step 2: Select "Species", "Database version", "Data type"**

Actinidia chinensis

Database version: release-47

Data type: ☒ Genome sequence (FASTA) ☐ Genome annotation (GTF)

Execute → **Step 3: Click this button to run the function**

What it does

This function can be used to retrieve **Genome sequences** and **Genome annotation** from Ensembl Plants (<https://plants.ensembl.org/index.html>), **RNA-Seq data** from NCBI (National Center for Biotechnology Information; <https://www.ncbi.nlm.nih.gov/>) GEO (Gene Expression Omnibus) or SRA (Short Read Archive) databases.

Inputs

Obtain Genome Sequences and Annotation is used to retrieve Genome sequences and Genome annotation.

Users can obtain genome sequences and annotation through selecting Species, Database version, and Data type.

Obtain RNA-Seq data in option Genome or RNA-Seq is used to retrieve RNA-Seq data.

easyMF provides two methods to automatically fetch RNA-Seq data using NCBI's tool *sra-toolkit* (version 2.3.5-2) for accession IDs or using wget for HTTP/FTP addresses.

When selecting **Fetch data through data ID**, Accession IDs of RNA-Seq data allows a list of accession IDs of RNA-Seq data.

When selecting **Fetch data through ftp address**, HTTP/FTP addresses of RNA-Seq data allows a list of HTTP/FTP addresses of RNA-Seq data.

Outputs

- Reference genome sequence
- Reference genome annotation
- Raw RNA-Seq data

History search datasets

Unnamed history 1 shown 2.02 GB

1: Reference genome sequence

format: data, database: Z

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

filter, lag

The following objects are masked from 'package:base':

intersect, setdiff, setequal, union

Loading required package: stringr

Loading required package

Click to view data

Click to download data

- The following screenshots show us how to download RNA-Seq data using easyMF.

Step 1: download test data provided by easyMF;

Step 2: upload test datum 01_data_ID in directory Test_data/01_Matrix_Preparation to history panel;

Galaxy Analyze Data Workflow Shared Data Visualization Help Login or Register Using 2.0 GB

Tools search tools

Get Data Upload File from your computer

easyMF

- I) MATRIX PREPARATION
- Data Retrieval
- Expression Matrix Generation
- Expression Quality Control
- II) MATRIX FACTORIZATION
- Matrix Factorization
- III) METAGENE-BASED DEEP MINING USING AM
- Functional Gene Discovery
- IV) METAGENE-BASED DEEP MINING USING PM
- Sample Clustering Analysis
- Temporal-spatial Analysis

Workflows

- All workflows

Download from web or upload from disk

Regular Composite Collection

Step 1: Click this button to upload data

Step 2: Choose the test datum provided by easyMF

01_Matrix_Preparation

01_data_ID 45 字节 文档 今天 下午 11:20

02_maize_referen...genome_chr10.fa 151 MB 文档 今天 下午 11:20

02_maize_referen...enome_chr10.gtf 56.4 MB 文档 今天 下午 11:20

02_RNA-Seq_data.tar.gz 160.9 MB gzip 压缩归档 今天 下午 11:20

03 sample information 62 字节 文档 今天 下午 11:20

Choose local file → **Step 3: Click this button to choose local file**

easyMF provides two methods to automatically fetch RNA-Seq data using NCBI's tool *sra-toolkit* (version 2.3.5-2) for accession IDs or using wget for HTTP/FTP addresses.

When selecting **Fetch data through data ID**, Accession IDs of RNA-Seq data allows a list of accession IDs of RNA-Seq data.

When selecting **Fetch data through ftp address**, HTTP/FTP addresses of RNA-Seq data allows a list of HTTP/FTP addresses of RNA-Seq data.

Outputs

- Reference genome sequence
- Reference genome annotation
- Raw RNA-Seq data

History search datasets

Unnamed history 1 shown 2.02 GB

1: Reference genome sequence

format: data, database: Z

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

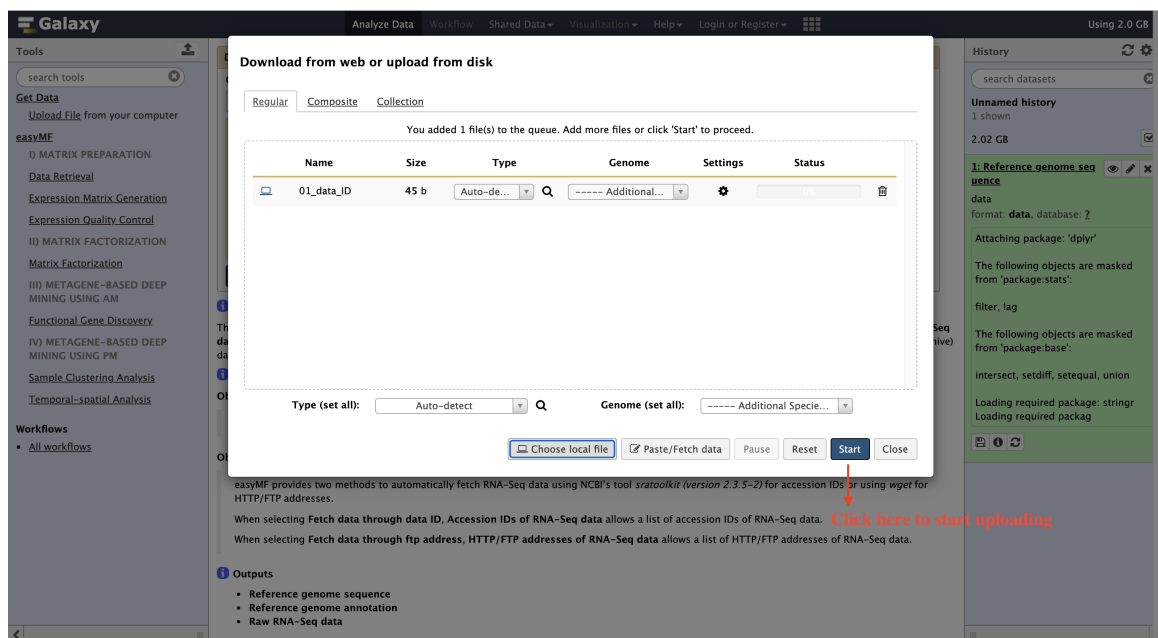
filter, lag

The following objects are masked from 'package:base':

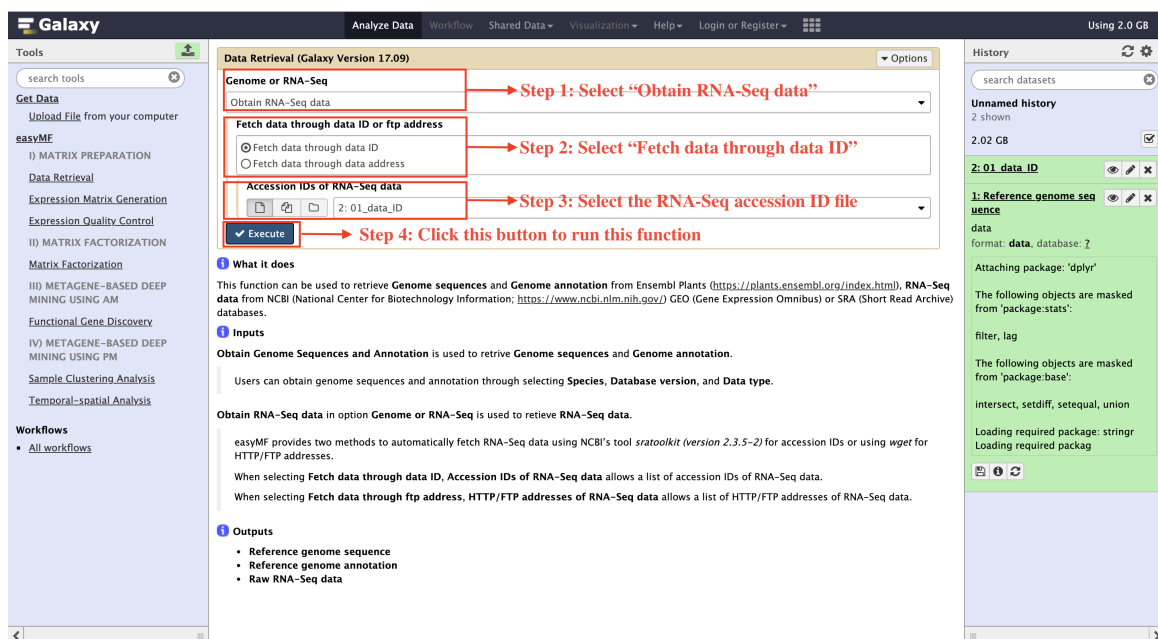
intersect, setdiff, setequal, union

Loading required package: stringr

Loading required package



Step 3: input the corresponding file, and run the function.



Running time

Running time for the test data depends on network speed.

2. Expression Matrix Generation

This function can be used to generate a gene expression matrix (genes in rows and individual samples in columns) through raw RNA-Seq quality control, read-genome alignment, and gene expression abundance calculation.

Inputs

In **Data** section

- **Reference genome sequence:** Reference genome sequence in FASTA format used for read-genome alignment.
- **Reference genome annotation:** Reference genome annotation in GTF format used to estimate gene expression abundance.
- **Raw RNA-Seq data:** A compressed file containing RNA-Seq data in tar.gz format.

In **Parameters** section, easyMF needs users set parameters used for "RNA-Seq quality control" and "Read-genome alignment".

For "RNA-Seq quality control"

- **Minimum read length:** A threshold of read length that reads shorter than the length will be discarded.
- **The quality value that a base is qualified:** A threshold of base quality value to trim low-quality reads.

For "Read-genome alignment"

- **Minimum intron length for RNA-Seq alignment**
- **Maximum intron length for RNA-Seq alignment**

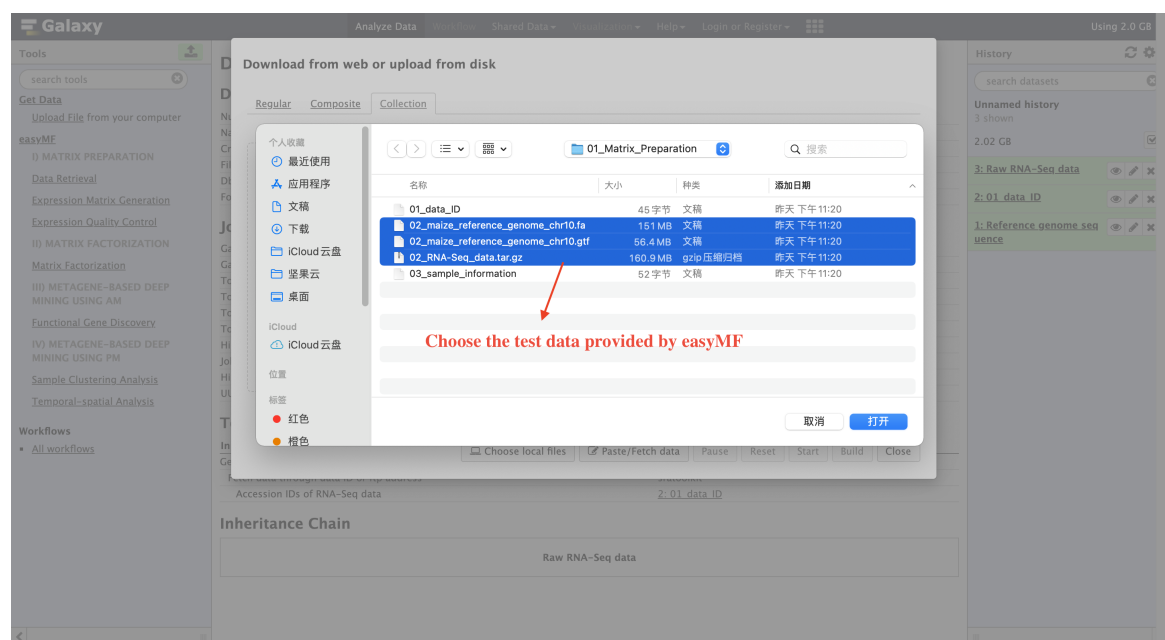
Output

- **Raw Gene Expression Matrix:** A gene expression matrix (genes in rows and individual samples in columns).

How to use this function

- Test data for this function are in directory `Test_data/01_Matrix_Preparation` including `02_maize_reference_genome_chr10.fa`, `02_maize_reference_genome_chr10.gtf`, and `02_RNA-Seq_data.tar.gz`.
- The following screenshots show us how to generate a gene expression matrix using easyMF.

Step 1: upload test data in directory `Test_data/01_Matrix_Preparation` to history panel;



Step 2: input the corresponding files and appropriate parameters, then run the function.

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Help | Login or Register | Using 8.0 GB

Tools | search tools | **Get Data** | Upload File from your computer | **easyMF** | I) MATRIX PREPARATION | Data Retrieval | **Expression Matrix Generation** | Expression Quality Control | II) MATRIX FACTORIZATION | Matrix Factorization | III) METAGENE-BASED DEEP MINING USING AM | Functional Gene Discovery | IV) METAGENE-BASED DEEP MINING USING PM | Sample Clustering Analysis | Temporal-spatial Analysis | **Workflows** | All workflows

Expression Matrix Generation (Galaxy Version 17.09) | Options

Data

- Reference genome sequence: 7: Raw Gene Expression Matrix
- Reference genome annotation: 7: Raw Gene Expression Matrix
- Raw RNA-Seq data: 7: Raw Gene Expression Matrix

Parameters

- RNA-Seq quality control**
 - Minimum read length: 15
 - The quality value that a base is qualified: 20
- Read-genome alignment**
 - Minimum intron length for RNA-Seq alignment: 20
 - Maximum intron length for RNA-Seq alignment: 10000
 - Threads: 3

Execute | **What it does** | **Inputs**

This function can be used to generate a gene expression matrix (genes in rows and individual samples in columns) through raw RNA-Seq quality control, read-genome alignment, and gene expression abundance calculation.

History | search datasets | Unnamed history | 7 shown | 7.99 GB

- 7: Raw Gene Expression Matrix
- 6: 02 RNA-Seq data.txt
- 5: 02 maize reference genome_chr10.gtf
- 4: 02 maize reference genome_chr10.fa
- 3: Raw RNA-Seq data
- 2: 01 data_ID
- 1: Reference genome s

Running time

This step will cost ~ 2 mins for the test data.

3. Expression Quality Control

Once gene expression matrix generated, to accurately implement MF-based analysis, quality of the gene expression matrix need to be improved, which can be operated through three different dimensions including **removing lowly expressed genes**, **removing outlier samples**, **removing batch effects**.

Inputs

For **Removing lowly expressed genes**

- Expression value of expressed genes:** Expression value of genes regarded as expressed.
- Minimum sample number:** The number of samples of expressed genes.

easyMF provides default values for these two parameters: **Expression value of expressed genes** (default as 1) and **Minimum sample number** (default as 3), which means genes regarded as expressed with expression value greater than 1 in at least 3 samples.

For **Removing outlier samples**

- Threshold of potential repeat samples:** Expression values between two samples are almost identical.
- Threshold of low-quality samples:** Sample distance between two RNA-Seq data.

For **Removing batch effects**

- Sample information:** RNA-Seq samples with batch information. In the text, the first column presents sample IDs, and the second column presents batch information distinguished with Arabic numerals.

```
SRR1765379 1
SRR1765380 1
SRR1765337 2
SRR1765338 2
```

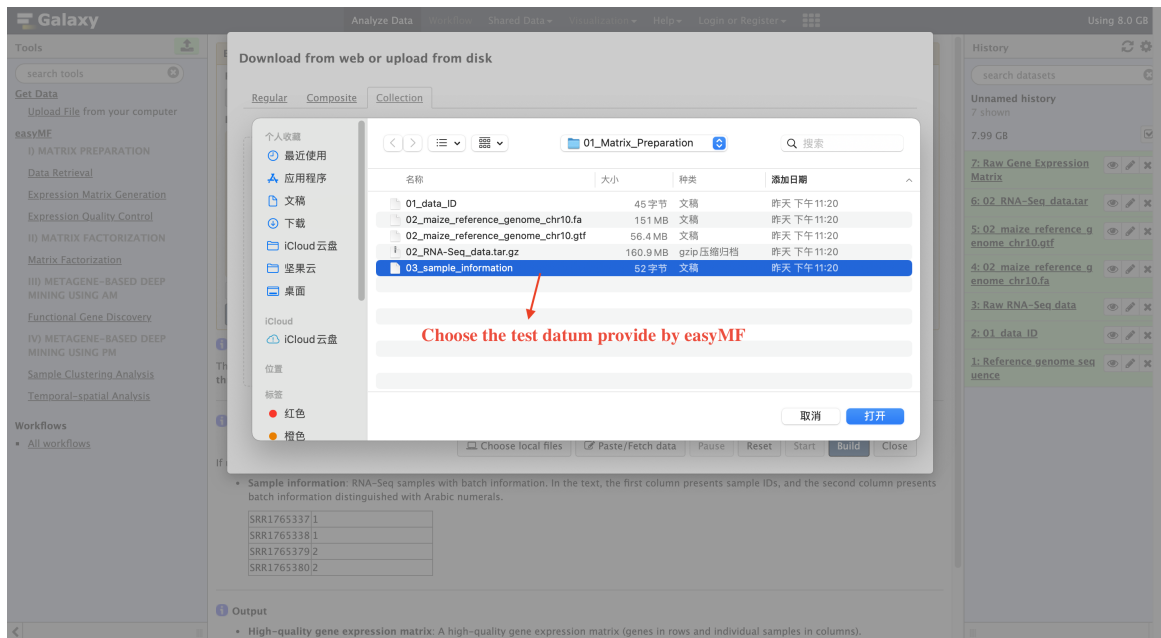
Output

- **High-quality gene expression matrix:** A high-quality gene expression matrix (genes in rows and individual samples in columns).

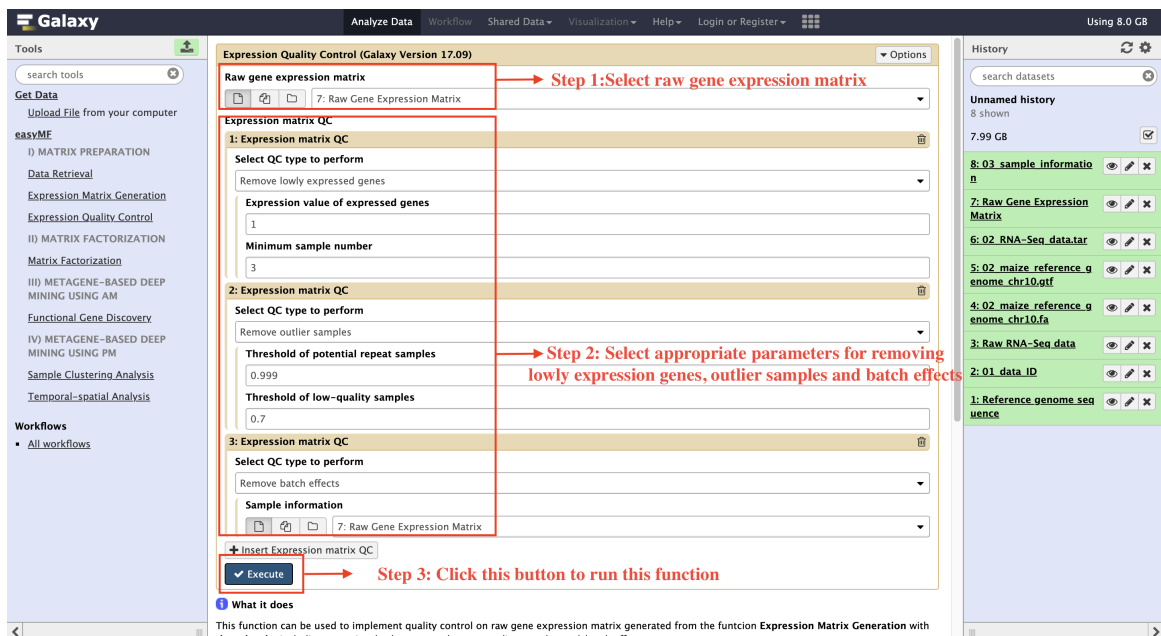
How to use this function

- Test data for this function are `03_sample_information` in `01_Matrix_Preparation` for **Sample information**, and **Raw gene expression matrix** generated by the function **Expression Matrix Generation**.
- The following screenshots show us how to generate a high-quality gene expression matrix using easyMF.

Step 1: upload test datum in directory `Test_data/01_Matrix_Preparation` to history panel;



Step 2: input the corresponding files and appropriate parameters, then run the function.



Running time

This step will cost ~ 10s for the test data.

