

mlPEA User Manual

(version 1.0)

- mlPEA is a user-friendly and multi-functionality platform specifically tailored to the needs of streamlined processing of m⁶A-Seq data in a reference genome-free manner. By taking advantage of machine learning (ML) algorithms, mlPEA enhanced the m⁶A-Seq data analysis by constructing robust computational models for identifying high-quality transcripts and high-confidence m⁶A-modified regions.
- mlPEA comprises four functional modules: **Data Preprocessing, Transcriptome Construction, m⁶A Calling, and Functional Exploration.**
- mlPEA was powered with an advanced packaging technology, which enables compatibility and portability.
- mlPEA project is hosted on <http://github.com/cma2015/mlPEA>
- mlPEA docker image is available at <http://hub.docker.com/r/malab/mlpea>

Data Preprocessing Module

This module provides four functions (see following table for details) to prepare epitranscriptome data.

Tools	Description	Input	Output	Time (test data)	Reference
Download File	Directly fetch epitranscriptome sequencing reads from NCBI's SRA database or other databases	SRR accession or HTTP/FTP link	Sequencing reads in SRA format	Depends on the network speed	SRA Toolkit
Sequence Data Preprocessing	Convert epitranscriptome sequencing reads from SRA to FASTQ format	Epitranscriptome sequencing reads in SRA format	Epitranscriptome sequencing reads in FASTQ format	~2 mins	SRA Toolkit
Assess Reads Quality	Check m ⁶ A-Seq reads quality and obtain high-quality reads	m ⁶ A-Seq reads in FASTQ format and adapter sequences in FASTA format	m ⁶ A-Seq reads in FASTQ format; MultiQC report in HTML	~2 mins	fastp ; MultiQC
Differential K-mer Analyses	Perform m ⁶ A-Seq differential k-mer analyses	m ⁶ A-Seq reads in FASTQ format	m ⁶ A-Seq reads in FASTA format	Depends on the file size	kmdiff

Download File

This function is designed to download epitranscriptome sequencing reads from NCBI SRA (Short Read Archive) database or from an user-specified HTTP/FTP link automatically. For the former, the **prefetch** function implemented in [SRA Toolkit](#) is wrapped to enable users to download sequencing data from NCBI SRA database; For the latter, **wget** command line is used to download the file according to an user-specified HTTP/FTP link.

Input

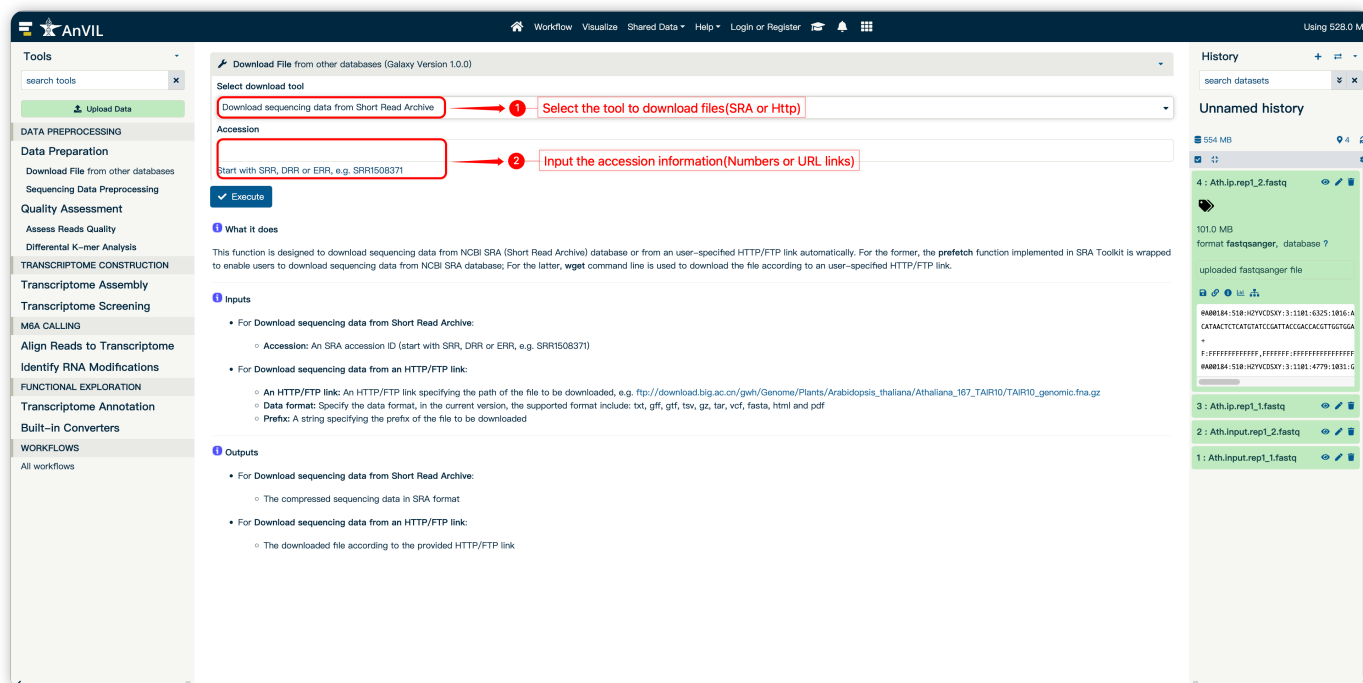
- For **Download sequencing data from Short Read Archive:**
 - **Accession:** An SRA accession ID (start with SRR, DRR or ERR, e.g. SRR1508371)
- For **Download sequencing data from an HTTP/FTP link:**
 - **An HTTP/FTP link:** An HTTP/FTP link specifying the path of the file to be downloaded, e.g. `ftp://download.big.ac.cn/gwh/Genome/Plants/Arabidopsis_thaliana/Athaliana_167_TAIR10/TAIR10_genomic.fna.gz`
 - **Data format:** Specify the data format, in the current version, the supported format include: txt, gff, gtf, tsv, gz, tar, vcf, fasta, html and pdf
 - **Prefix:** A string specifying the prefix of the file to be downloaded

Output

- For **Download sequencing data from Short Read Archive:**
 - The compressed sequencing data in SRA format
- For **Download sequencing data from an HTTP/FTP link:**
 - The downloaded file according to the provided HTTP/FTP link

How to use this function

- The following screenshot shows us how to download sequencing reads in SRA format



Sequence Data Preprocessing

This function wrapped **fastq-dump** function implemented in SRA Toolkit. See <http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software> for details.

Input

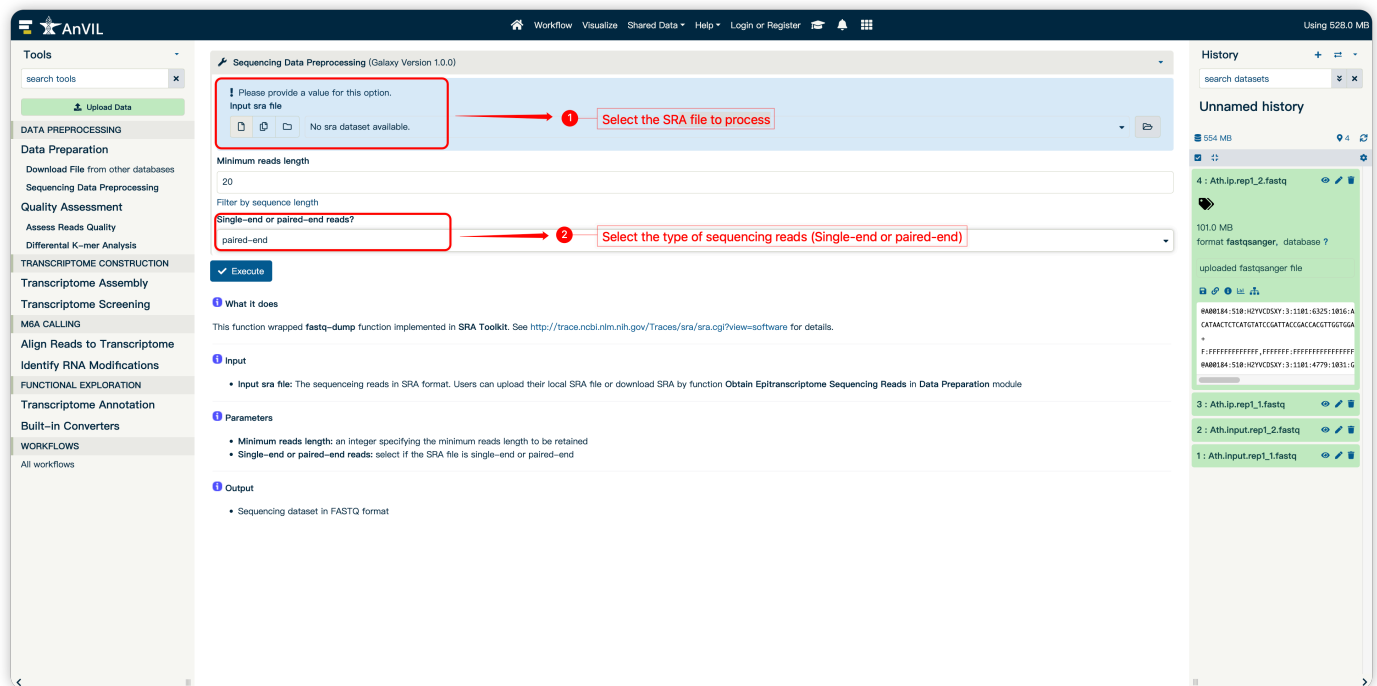
- **Input sra file:** The sequencing reads in SRA format. Users can upload their local SRA file or download SRA by function **Download File** in **Data Preparation** module

Output

- Sequencing dataset in FASTQ format

How to use this function

- The following screenshot shows us how to use this function to convert sequencing reads in SRA format to FASTQ format



Assess Reads Quality

In this function, two existing NGS tools MultiQC and fastp are integrated to check sequencing reads quality and obtain high-quality reads, respectively.

Input

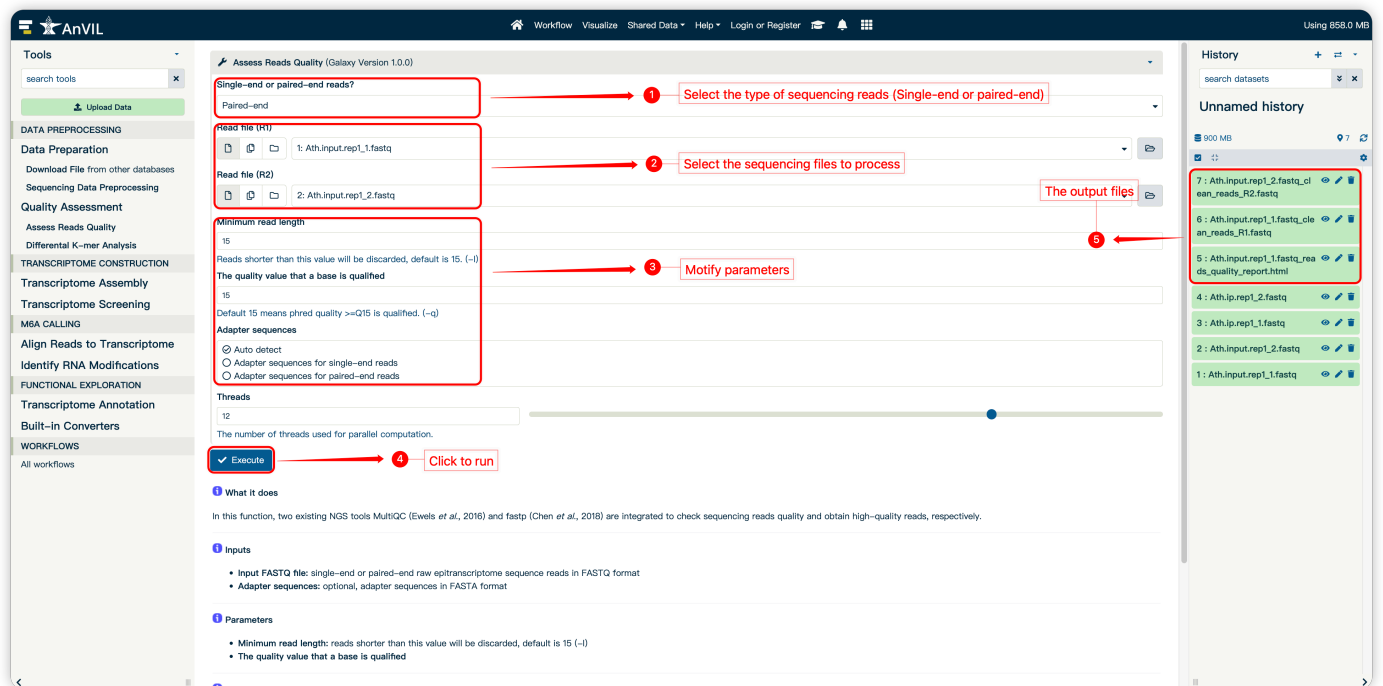
- **Input FASTQ file:** single-end or paired-end raw epitranscriptome sequence reads in FASTQ format
- **Adapter sequences:** optional, adapter sequences in FASTA format

Output

- Clean reads in FASTQ format
- Clean reads MultiQC report in HTML format

How to use this function

- The following screenshot shows us how to use this function to check m⁶A-Seq reads quality and obtain high-quality reads.



Differential *K*-mer Analyses

In this function, one existing disk-based program kmdiff is integrated to count *k*-mers from (possibly gzipped) FASTQ/FASTA files.

Input

- **Input FASTQ file:** IP and input cleaned epitranscriptome sequence reads in FASTQ format

Output

- **Differential *k*-mer sequences (input sample) in FASTA format**
- **Differential *k*-mer sequences (IP sample) in FASTA format**

How to use this function

- The following screenshot shows us how to use this function.

AnVIL

Workflow

Visualize

Shared Data

Help

Login or Register

Using 1.0 GB

Tools

search tools

Upload Data

DATA PREPROCESSING

Data Preparation

Download File from other databases

Sequencing Data Preprocessing

Quality Assessment

Assess Reads Quality

Differential K-mer Analysis

TRANSCRIPTOME CONSTRUCTION

Transcriptome Assembly

Transcriptome Screening

MSA CALLING

Align Reads to Transcriptome

Identify RNA Modifications

FUNCTIONAL EXPLORATION

Transcriptome Annotation

Built-in Converters

WORKFLOWS

All workflows

Differential K-mer Analysis (Galaxy Version 1.0.0)

Single-end or paired-end reads?

paired-end

The FASTQ file(R1) of Input

6: Ath.input.rep1_1.fastq_clean_reads_R1.fastq

The FASTQ file(R2) of Input

10: Ath.ip.rep1_2.fastq_clean_reads_R2.fastq

The FASTQ file(R1) of IP

9: Ath.ip.rep1_1.fastq_clean_reads_R1.fastq

The FASTQ file(R2) of IP

10: Ath.ip.rep1_2.fastq_clean_reads_R2.fastq

k-mers length

8

k-mer length (k from 8 to 20; default: 8), (-k)

hard min

20

min abundance to keep a k-mer. (---hard-min)

Threads

11

The number of threads used for parallel computation.

Execute

What it does

In this function, one existing disk-based program kmdiff is integrated to count k-mers from (possibly gzipped) FASTQ/FASTA files.

Inputs

Input FASTQ file: IP and input cleaned epitranscriptome sequence reads in FASTQ format

Parameters

k-mers length: k-mer length (k from 8 to 20; default: 8)

hard-min: min abundance to keep a k-mer

1 Select the type of sequencing reads(Single-end or paired-end)

2 Select the clean reads to process

3 Modify parameters

4 Click to run

5 The output files

History

search datasets

Unnamed history

1.12 GB

12

12 : Differential K-mer sequences(IP sample)

11 : Differential K-mer sequences(input sample)

10 : Ath.ip.rep1_2.fastq_clean_reads_R2.fastq

9 : Ath.ip.rep1_1.fastq_clean_reads_R1.fastq

8 : Ath.ip.rep1_1.fastq_reads_quality_report.html

7 : Ath.input.rep1_2.fastq_clean_reads_R2.fastq

6 : Ath.input.rep1_1.fastq_clean_reads_R1.fastq

5 : Ath.input.rep1_1.fastq_reads_quality_report.html

4 : Ath.ip.rep1_2.fastq

3 : Ath.ip.rep1_1.fastq

2 : Ath.input.rep1_2.fastq

1 : Ath.input.rep1_1.fastq