mlPEA User Manual

(version 1.0)

- mlPEA is a user-friendly and multi-functionality platform specifically tailored to the needs of streamlined processing of m^{6}A-Seq data in a reference genome-free manner. By taking advantage of machine learning (ML) algorithms, mlPEA enhanced the m^{6}A-Seq data analysis by constructing robust computational models for identifying high-quality transcripts and high-confidence m^{6}A-modified regions.
- mlPEA comprises four functional modules: **Data Preprocessing, Transcriptome Construction, m^{6}A 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

m^{6}A Calling Module

This module provides step-by-step functions required for epitranscriptome reads mapping and identification of RNA modifications.

Align Reads to Assembled Transcriptome

Bowtie2 is wrapped to align epitranscriptome reads to assembled transcriptome.

Tools	Description	Input	Output	Time (test data)	Reference
Bowtie2	Bowtie2 is a short read aligner which achieves a combination of high speed, sensitivity and accuracy by combining the strengths of the full-text minute index with the flexibility and speed of hardware-accelerated dynamic programming algorithms, therefore bowtie2 is suitable for large genomes	Epitranscriptome sequencing reads in FASTQ format and assembled transcriptome sequences in FASTA format	Read alignments in SAM/BAM format	~15 mins	(Grabherr *et al.*, 2011)

Identify RNA Modifications

Identify RNA Modifications implements three pipelines.

Tools	Description	Input	Output	Time (test data)	Reference
Peak Calling	Using the SlidingWindow algorithm to identify regions with statistically significant enrichment of m^6A signals compared to the background.	Assembled transcripts in FASTA format; Reads coverage file in IP and input sample; Total mapped reads number in IP and input sample	The enriched peak region matrix in BED format	~10 mins	(Zhai *et al*, 2018)
Merge Biological Replicates	Obtain consistent RNA modifications among multiple biological replicates.	Peak regions for biological replicates	Consistent peak regions among multiple biological replicates in BED format	~1 mins	in-house scripts
Machine Learning-based Peak Screening	All peaks derived from the assembled transcripts are designated as positive samples and utilized as input for the weakly supervised learning framework to facilitate model training.	Assembled transcripts in FASTA format; The enriched peak region matrix in BED format	High condfidence Peak region in BED format	~6mins	(Huang *et al*, 2021; Song *et al*, 2024)

Align Reads to Assembled Transcriptome

In this function, clean reads are aligned to high-quality assembled transcripts using **Bowtie2**, with the *de novo* assembled transcriptome serving as the reference.

Input

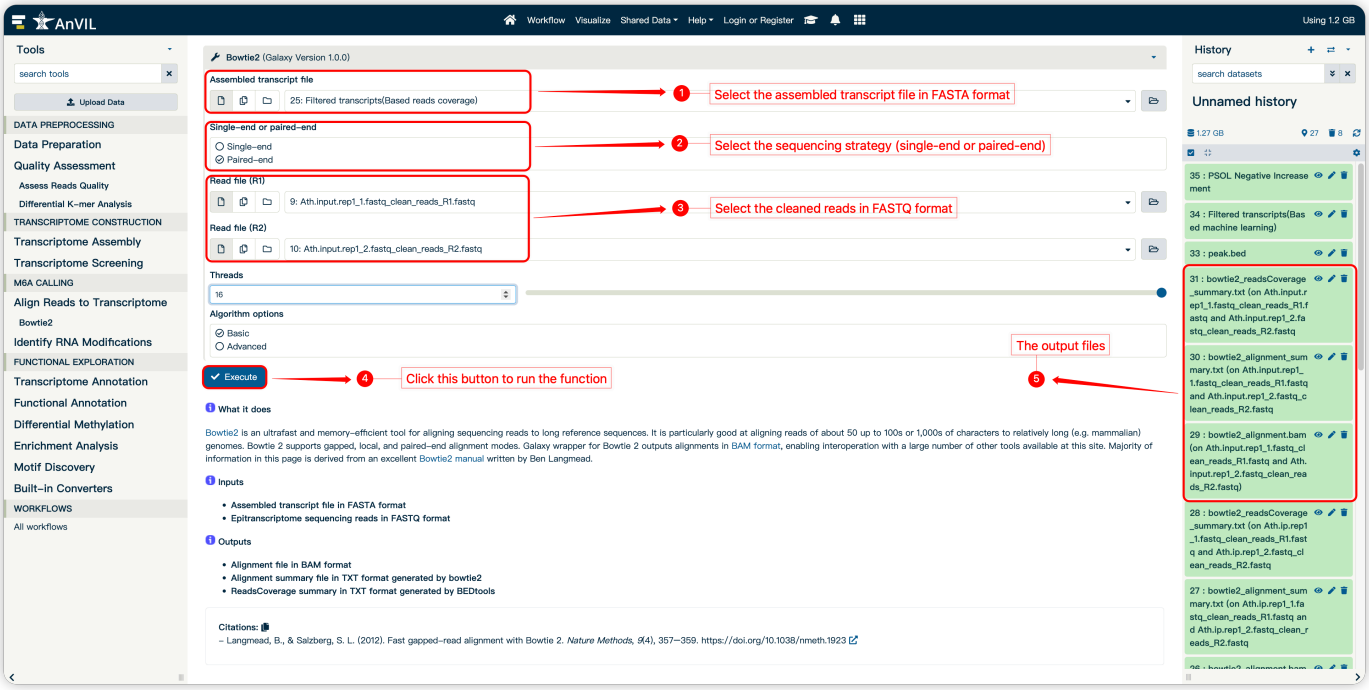
- **Epitranscriptome sequencing reads in FASTQ format**
- **Assembled transcripts in FASTA format**

Output

- **Alignment reaults in BAM format**
- **Alignment summary file generated by bowtie2**
- **Reads coverage summary in TXT format generated by BEDtools**

How to use this function

- The following screenshot shows us how to use this function



Peak Calling

In this function, peak calling was performed using **PEA**, which used the SlidingWindow algorithm to identify regions with statistically significant enrichment of m^6A signals compared to the background.

Input

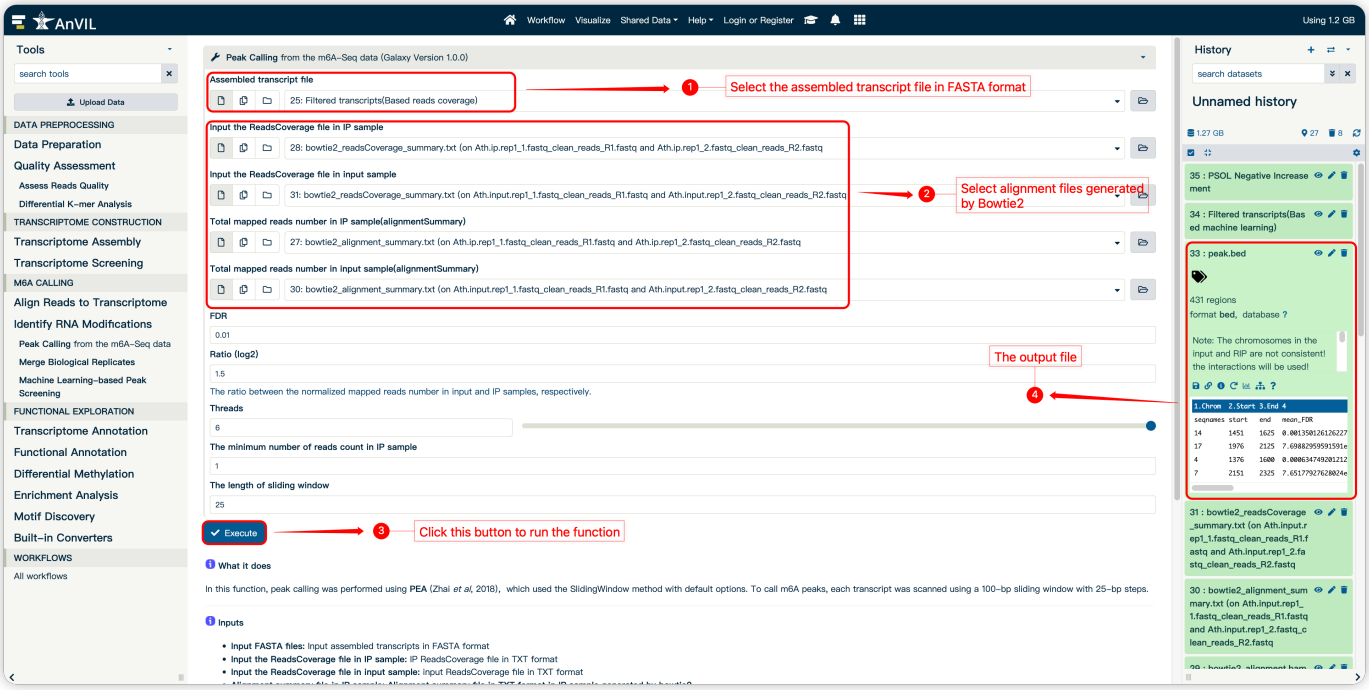
- **Input the assembled transcript file:** assembled transcripts in FASTA format
- **Input the reads coverage files:** Reads coverage file in IP and input sample in TXT format
- **Input the alignment summary file:** Total mapped reads number in IP and input sample in TXT format

Output

- **The enriched peak region matrix in BED format**

How to use this function

- The following screenshot shows us how to use this function



Merge Biological Replicates

In this function, mIPEA obtained consistent RNA modifications among multiple biological replicates.

Input

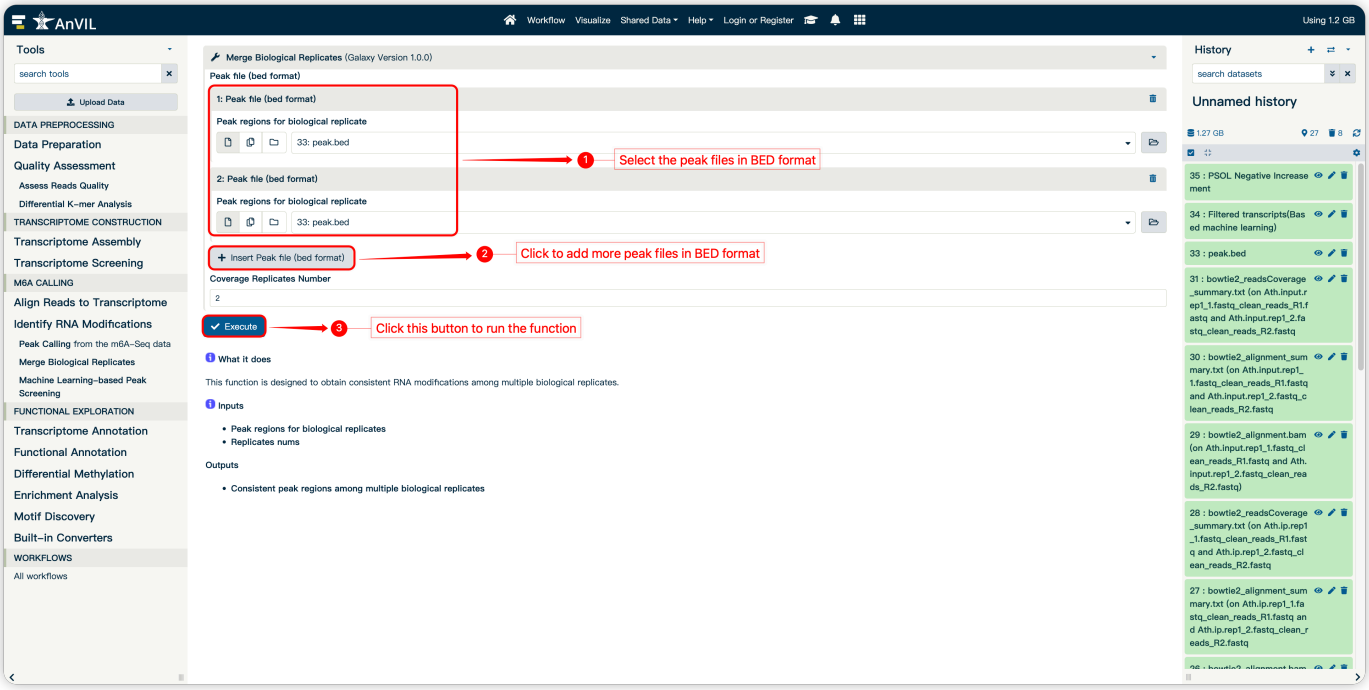
- **Peak regions for biological replicates**

Output

- **Consistent peak regions among multiple biological replicates in BED format**

How to use this function

- The following screenshot shows us how to use this function



Machine Learning-based Peak Screening

All peaks derived from the assembled transcripts are designated as positive samples and utilized as input for the weakly supervised learning framework to facilitate model training.

Input

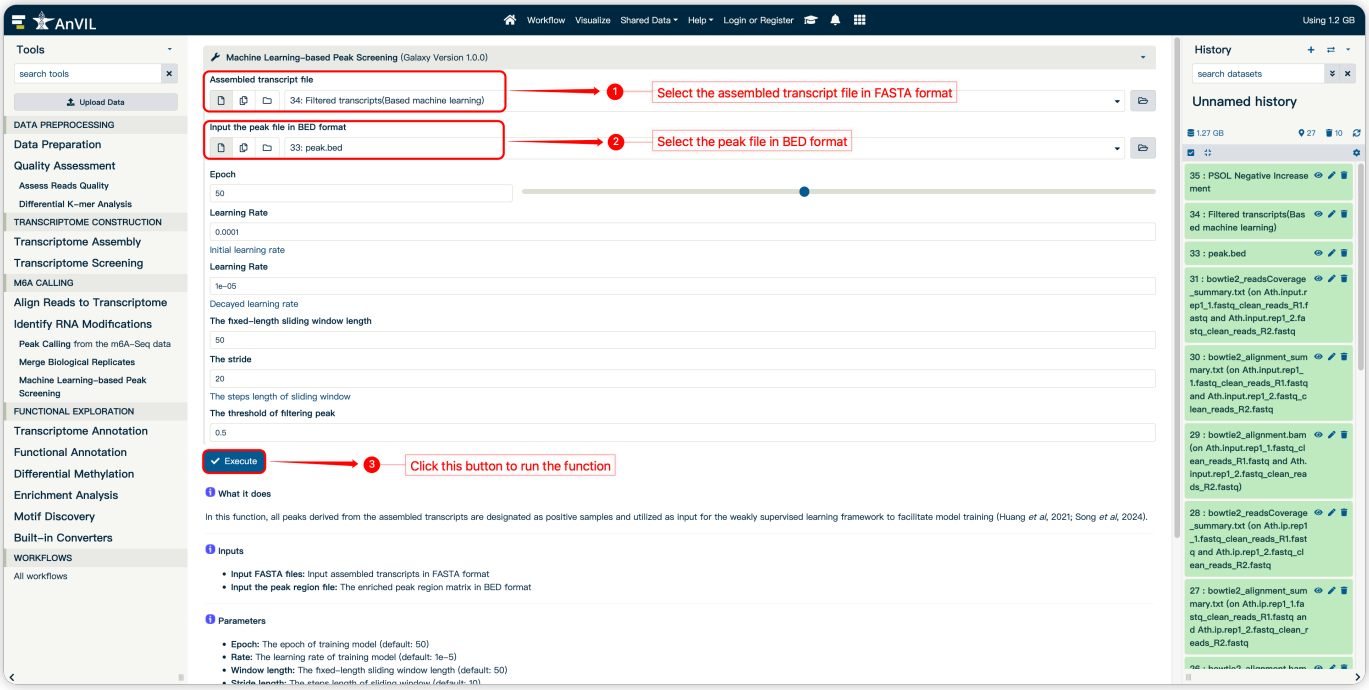
- **Input the assembled transcripts file:** assembled transcripts in FASTA format
- **Input the peak region file:** The enriched peak region matrix in BED format

Output

- **High condfidence Peak region in BED 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 1.2 GB
Tools
search tools
Upload Data
DATA PREPROCESSING
Data Preparation
Quality Assessment
Assess Reads Quality
Differential K-mer Analysis
TRANSCRIPTOME CONSTRUCTION
Transcriptome Assembly
Transcriptome Screening
M6A CALLING
Align Reads to Transcriptome
Identify RNA Modifications
Peak Calling from the m6A-Seq data
Merge Biological Replicates
Machine Learning-based Peak Screening
FUNCTIONAL EXPLORATION
Transcriptome Annotation
Functional Annotation
Differential Methylation
Enrichment Analysis
Motif Discovery
Built-in Converters
WORKFLOWS
All workflows
Machine Learning-based Peak Screening (Galaxy Version 1.0.0)
Assembled transcript file
34: Filtered transcripts(Based machine learning)
1
Select the assembled transcript file in FASTA format
Input the peak file in BED format
33: peak.bed
2
Select the peak file in BED format
Epoch
50
Learning Rate
0.0001
Initial learning rate
Learning Rate
1e-05
Decayed learning rate
The fixed-length sliding window length
50
The stride
20
The steps length of sliding window
The threshold of filtering peak
0.5
Execute
3
Click this button to run the function
What it does
In this function, all peaks derived from the assembled transcripts are designated as positive samples and utilized as input for the weakly supervised learning framework to facilitate model training (Huang et al, 2021; Song et al, 2024).
Inputs
Input FASTA files: Input assembled transcripts in FASTA format
Input the peak region file: The enriched peak region matrix in BED format
Parameters
Epoch: The epoch of training model (default: 50)
Rate: The learning rate of training model (default: 1e-5)
Window length: The fixed-length sliding window length (default: 50)
Stride length: The steps length of sliding window (default: 10)
History
search datasets
Unnamed history
1.27 GB
27
10
35 : PSOL Negative Increasement
34 : Filtered transcripts(Based machine learning)
33 : peak.bed
31 : bowtie2_readsCoverage_summary.txt (on Ath.input.rep1_1.fastq_clean_reads_R1.fastq and Ath.input.rep1_2.fastq_clean_reads_R2.fastq
30 : bowtie2_alignment_summary.txt (on Ath.input.rep1_1.fastq_clean_reads_R1.fastq and Ath.input.rep1_2.fastq_clean_reads_R2.fastq
29 : bowtie2_alignment.bam (on Ath.input.rep1_1.fastq_clean_reads_R1.fastq and Ath.input.rep1_2.fastq_clean_reads_R2.fastq)
28 : bowtie2_readsCoverage_summary.txt (on Ath.ip.rep1_1.fastq_clean_reads_R1.fastq and Ath.ip.rep1_2.fastq_clean_reads_R2.fastq
27 : bowtie2_alignment_summary.txt (on Ath.ip.rep1_1.fastq_clean_reads_R1.fastq and Ath.ip.rep1_2.fastq_clean_reads_R2.fastq