

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>

m⁶A 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 <i>et al.</i> , 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 ⁶ A 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 <i>et al</i> , 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 |
| ML-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 confidence Peak region in BED format | ~6mins | (Huang <i>et al</i> , 2021; Song <i>et al</i> , 2024) |

Align Reads to Assembled Transcriptome

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

Input

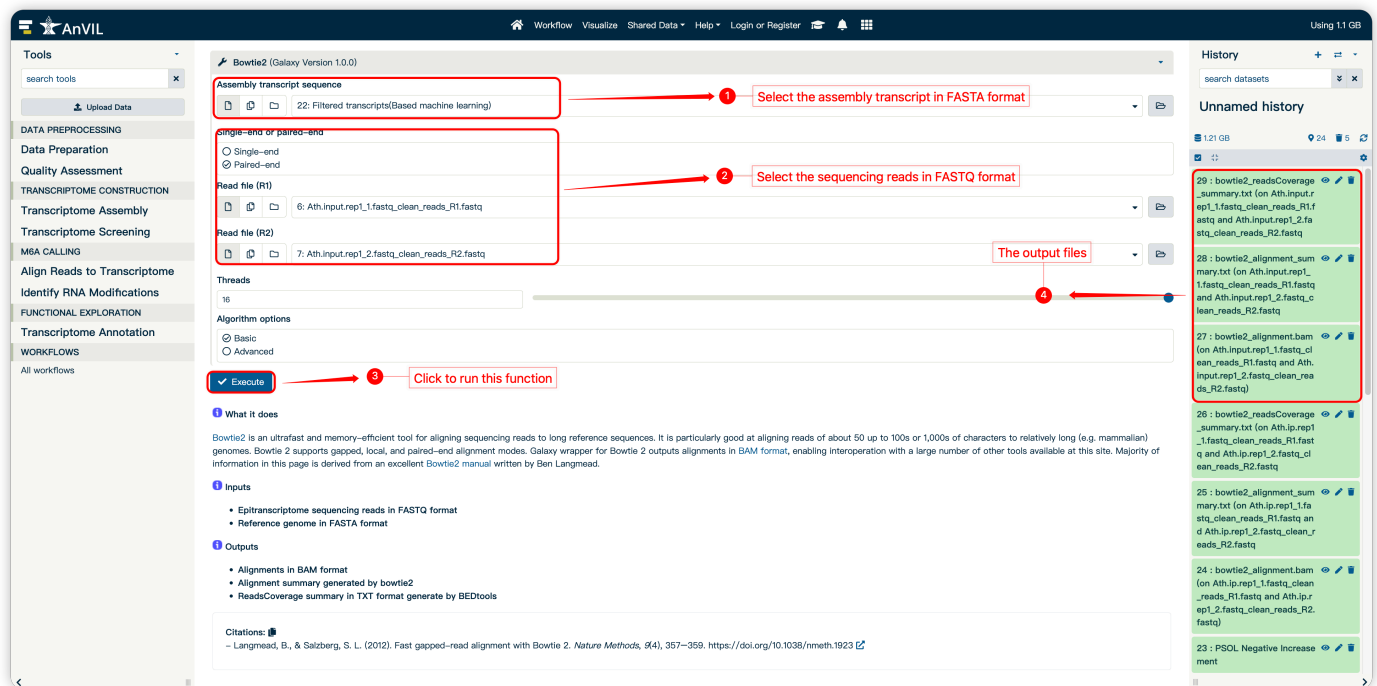
- Epitranscriptome sequencing reads in FASTQ format
- Assembled transcripts in FASTA format

Output

- Alignment results in BAM format
- Alignment summary generated by bowtie2
- Reads coverage summary in TXT format generate 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 m⁶A signals compared to the background.

Input

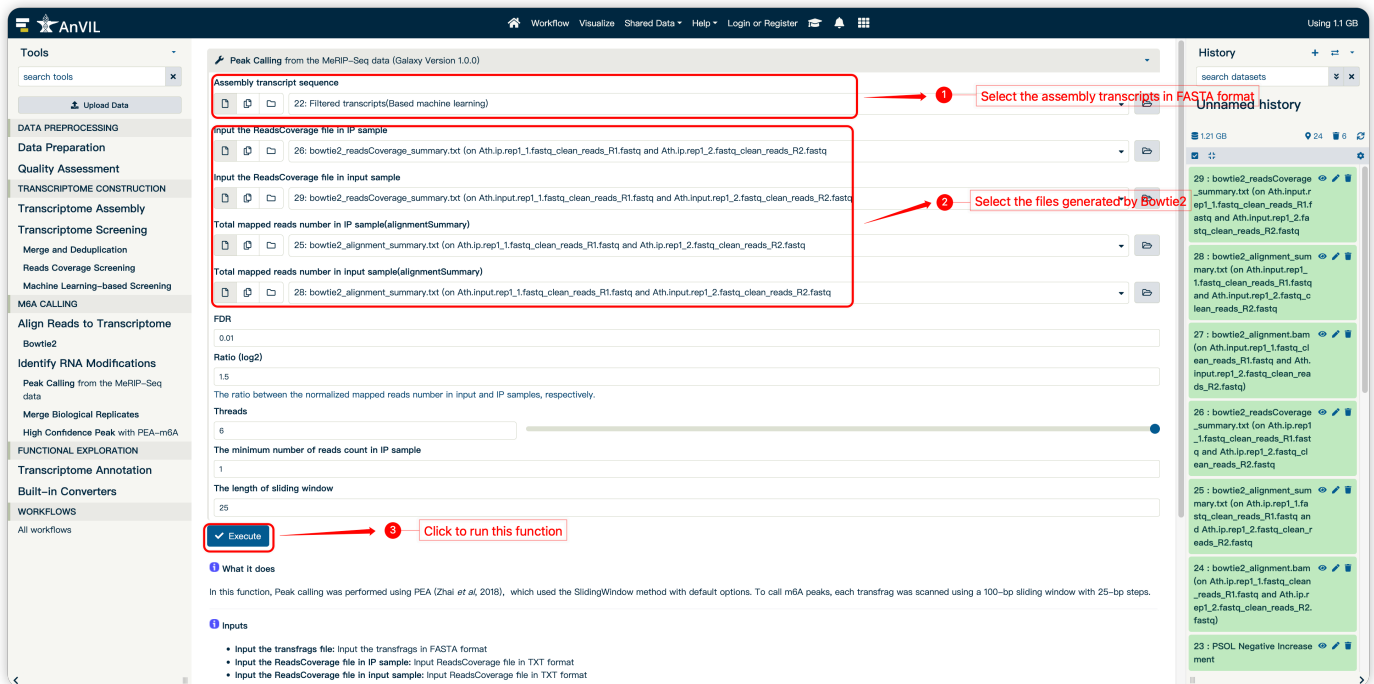
- **Input the assembled transcripts file:** assembled transcripts in FASTA format
- **Input the reads coverage files:** Reads coverage file in IP and input sample in TXT format
- **Input the alignment summary file:** Total mapped reads number in IP and input sample in TXT format

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, mlPEA obtained consistent RNA modifications among multiple biological replicates.

Input

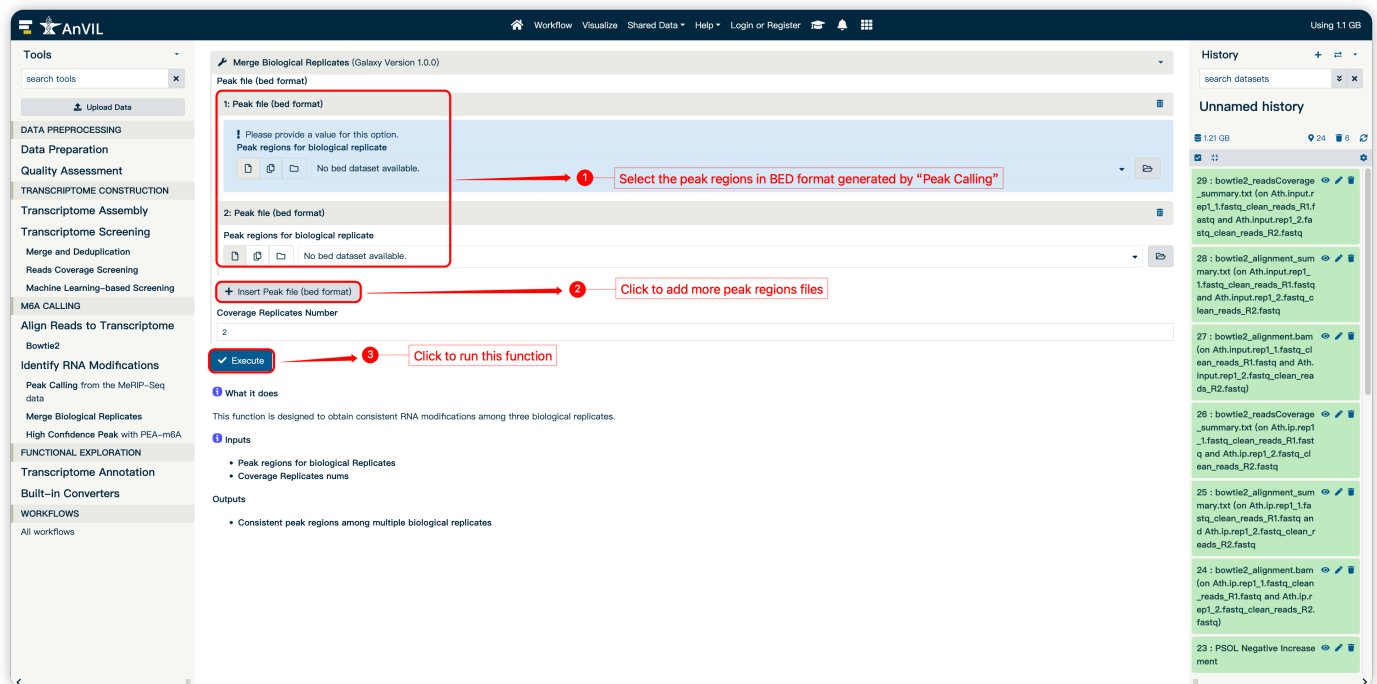
- Peak regions for biological replicates

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



ML-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 confidence Peak region in BED format**

How to use this function

- The following screenshot shows us how to use this function

[illegible]