

# 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<sup>6</sup>A-Seq data in a reference genome-free manner. By taking advantage of machine learning (ML) algorithms, mlPEA enhanced the m<sup>6</sup>A-Seq data analysis by constructing robust computational models for identifying high-quality transcripts and high-confidence m<sup>6</sup>A-modified regions.
- mlPEA comprises four functional modules: **Data Preprocessing, Transcriptome Construction, m<sup>6</sup>A Calling, and Functional Exploration.**
- mlPEA was powered with an advanced packaging technology, which enables compatibility and portability.
- mlPEA project is hosted on <https://github.com/cma2015/mlPEA>
- mlPEA Docker image is available at <https://hub.docker.com/r/malab/mlpea>

## Functional Exploration

This module provided five functions to perform functional exploration of m<sup>6</sup>A-Seq data

Functions	Description	Input	Output	Time (test data)	Reference
<b>ML-based Transcript Annotation</b>	Predict the coding region of assembled transcripts	Assembled transcripts in FASTA format	Prediction scores of translation initiation and termination sites of assembled transcripts in txt format	~3 min	In-house scripts
<b>Function Annotation</b>	Annotate the functions of coding transcripts	RNA modifications in BED format and assembled transcripts in FASTA format	Functions corresponding to transcripts in txt format	~3 min	
<b>Differential Methylation</b>	Identify differential methylation modifications under multiple conditions	Assembled transcripts in FASTA format and alignment file in bam format	RNA differential modifications in BED format	~5 min	In-house scripts
<b>Enrichment Analysis</b>	Perform GO or KEGG enrichment analysis for any species	Transcript list and function annotation results	The enriched GO/KEGG terms	~3 min	
<b>Motif Discovery</b>	Integrate MEME-ChIP and HOMER to performed <i>de novo</i> motif discovery	RNA modifications in BED format and assembled transcriptome sequences in FASTA format	Discovered motifs in HTML format	~1 min	

## ML-based Transcript Annotation

In this function, we utilized **TranslationAI**, a deep neural network to directly predict and analyze translation initiation (TIS) and termination sites (TSS) from transcripts.

## Input

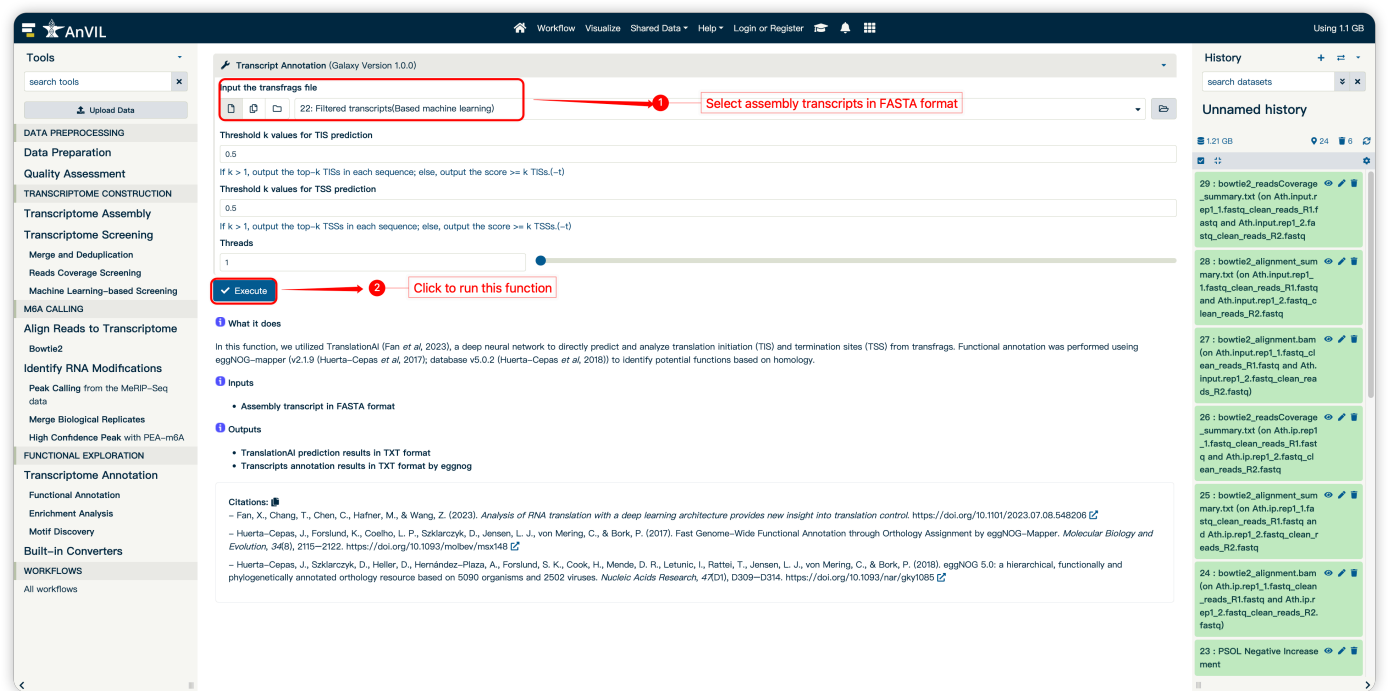
- Assembly transcript in FASTA format

## Output

- TranslationAI prediction results in TXT format
- Transcripts annotation results in TXT format

## How to use this function

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



## Function Annotation

In this function, functional annotation was performed using eggNOG-mapper (database v5.0.2) to identify potential functions based on homology.

## Input

- Assembled transcript in FASTA format

## Output

- Assembled transcripts function annotation results in TXT format by eggNOG

## How to use this function

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

# Differential Methylation

---

In this function, for pair-wised m<sup>6</sup>A-Seq data, mlPEA can identify differentially methylated regions (DMRs) using **QNB**, with the negative-binomial distribution model to capture the within-group variability of m<sup>6</sup>A methylation level across all samples.

## Input

- **Assembled transcript in FASTA format**
- **Alignment results in BAM format**

## Output

- **The differential peak region matrix in TXT format**

## How to use this function

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

# Enrichment Analysis

---

In this function, for a set of m<sup>6</sup>A-modified transcripts of interest, Kyoto Encyclopedia of Genes and Genomes (KEGG) and gene ontology (GO) enrichment analysis is performed utilizing the R package **clusterProfiler**.

## Input

- **Assembled transcripts function annotation results in TXT format by eggNOG**
- **RNA modifications transcript list**

## Output

- The enriched GO/KEGG terms
- A PDF focument of top enriched GO/KEGG terms

## How to use this function

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

# Motif Discovery

---

This function integrates MEME-ChIP and HOMER to perform *de novo* motif discovery.

## Input

- **RNA modifications regions in BED format**

- Assembled transcript in FASTA format

## Output

- An HTML report generated by MEME-ChIP or HOMER

## How to use this function

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

The screenshot displays the AnVIL Motif Discovery (Galaxy Version 1.0.0) interface. The left sidebar shows a navigation menu with categories like Tools, DATA PREPROCESSING, Data Preparation, Quality Assessment, TRANSCRIPTOME CONSTRUCTION, Transcriptome Assembly, Transcriptome Screening, Reads Coverage Screening, Machine Learning-based Screening, MSA CALLING, Align Reads to Transcriptome, Bowtie2, Identify RNA Modifications, Peak Calling from the MeRIP-Seq data, Merge Biological Replicates, High Confidence Peak with PEA-m6A, FUNCTIONAL EXPLORATION, Transcriptome Annotation, Functional Annotation, Enrichment Analysis, Motif Discovery, Built-in Converters, and WORKFLOWS. The main panel is titled 'Motif Discovery (Galaxy Version 1.0.0)' and contains several sections: 'Please provide a value for this option. RNA modifications (peak regions or single nucleotide resolution) in BED format' with a file upload button and a message 'No bed dataset available.'; 'Assembled transcript sequences in FASTA format' with a file upload button and a message '22: Filtered transcripts(Based machine learning)'; 'Select a method' with three radio buttons: 'HOMER: Hypergeometric Optimization of Motif Enrichment', 'MEME-ChIP: motif discovery, enrichment analysis and clustering on large nucleotide datasets', and 'DREME: Discriminative Regular Expression Motif Elicitation'; 'Options Configuration' with a dropdown menu set to 'Basic'; and an 'Execute' button. The bottom section includes 'What it does', 'Inputs', 'Outputs', and 'Citations'. Four red arrows with numbers 1 through 4 point to specific elements: 1 points to the 'Please provide a value for this option...' section, 2 points to the 'Assembled transcript sequences in FASTA format' section, 3 points to the 'Select a method' section, and 4 points to the 'Execute' button. The right sidebar shows a 'History' section with a search bar and a list of datasets.

1. Select the peak region file in BED format

2. Select the assembly transcripts in FASTA format

3. Select the method to use

4. Click to run this function