

0. Introduction

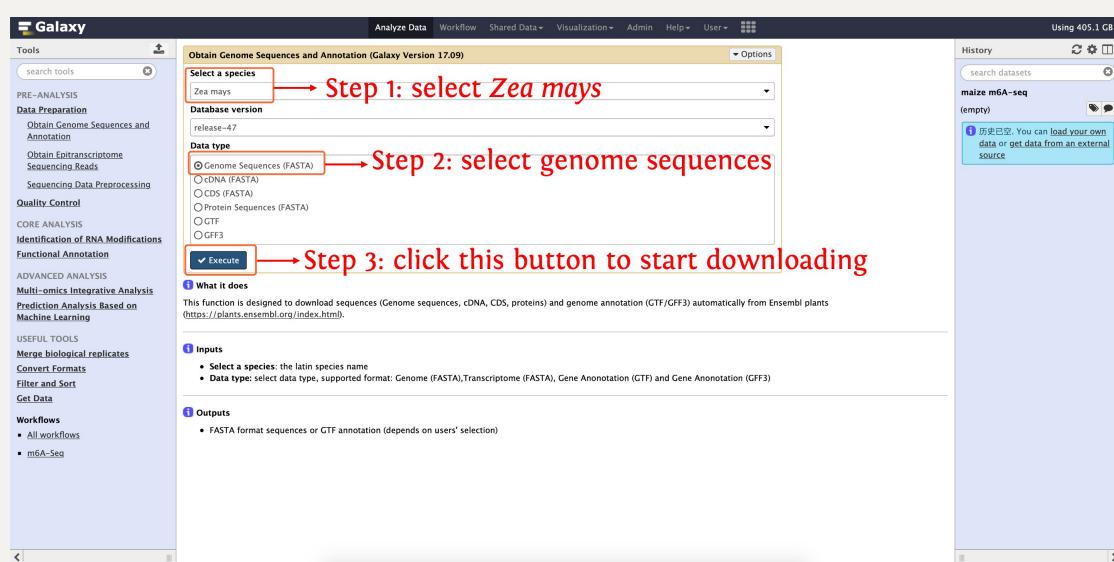
In this tutorial, we will show how to use deepEA (<https://deepea.nwafu.edu.cn> or <http://3.9.101.176.205:4006>) to perform comprehensive m⁶A sequencing data analysis. The m⁶A - immunoprecipitated (IP) and input (non-IP) samples with two biological replicates were extracted from the 14-day-old seedlings of maize (*Zea mays* L.) inbred lines B73, and then sequenced with high-throughput sequencing technology. The SRA accessions of four datasets are listed in the following table. The More information regarding these maize m⁶A sequencing datasets is available in the reference ([Luo et al., 2020](#)).

| SAMPLES | EXPERIMENTS | REPLICATES |
|------------|-------------|-------------|
| SRR8383013 | IP | Replicate 1 |
| SRR8383014 | IP | Replicate 2 |
| SRR8383017 | input | Replicate 1 |
| SRR8383018 | input | Replicate 2 |

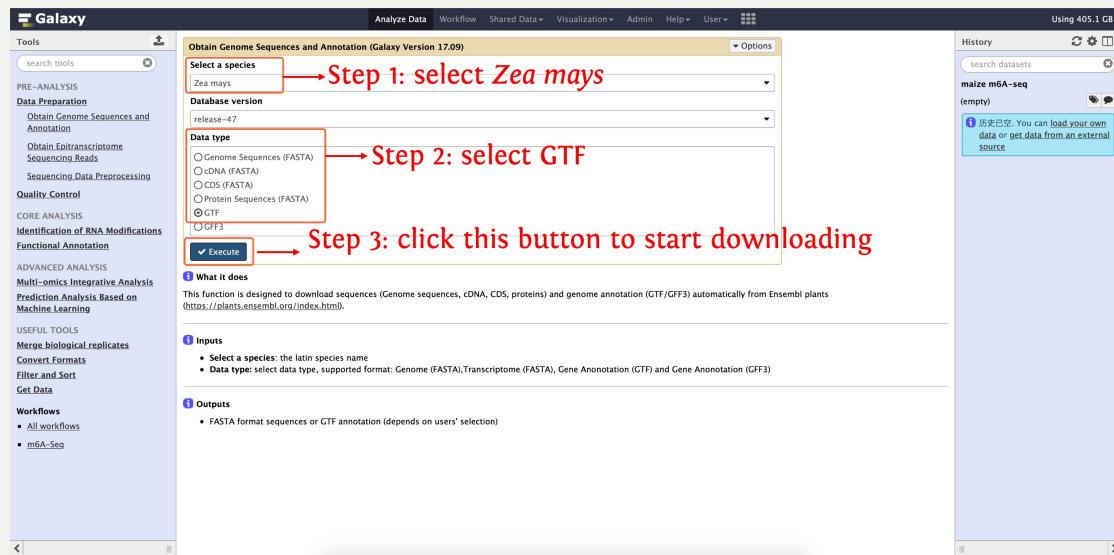
1. Download *Zea mays* reference genome sequences and annotation

Before analyzing m⁶A sequencing data, we firstly use the function **Obtain Genome Sequences and Annotation** in **Data Preparation** module to download *Zea mays* B73 reference genome sequences and GTF annotation, the following two screenshots shows details about how to execute this step:

Step 1: download *Zea mays* reference genome sequences in FASTA format



Step 2: download *Zea mays* genome annotation file in GTF format

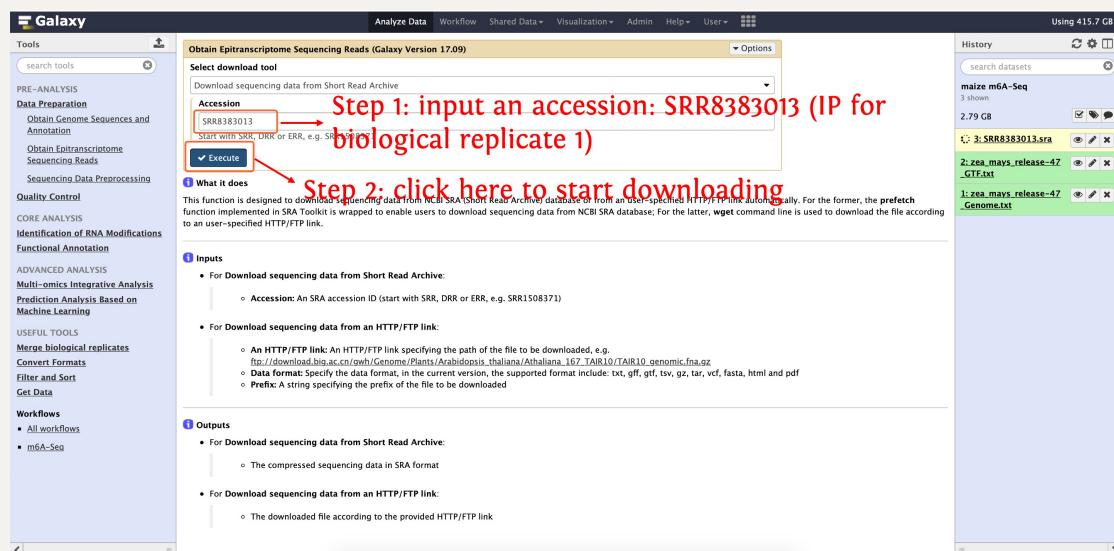


After that, reference genome sequences (named as **zea_mays_release-47_Genome.txt**) and annotation (named as **zea_mays_release-47_GTF.txt**) will be listed in your History Panel.

2. Download raw m⁶A sequencing reads

After finishing downloading *Zea mays* B73 reference genome sequences and annotation, we start to download raw m⁶A sequencing reads from NCBI SRA's database, this process can be finished by the function **Obtain Epitranscriptome Sequencing Reads** in Data Preparation module. Please see the following screenshots for details:

Step 1: download IP sample for biological replicate 1 (SRR8383013)



Step 2: download IP sample for biological replicate 2 (SRR8383014)

Step 1: input an accession: SRR8383014 (IP for biological replicate 2)

Step 2: click here to start downloading

Step 3: download input sample for biological replicate 1 (SRR8383017)

Step 1: input an accession: SRR8383017 (input for biological replicate 1)

Step 2: click here to start downloading

Step 4: download input sample for biological replicate 2 (SRR8383018)

Step 1: input an accession: SRR8383018 (input for biological replicate 2)

Step 2: click here to start downloading

The History panel on the right shows the download progress and results:

- 6: SRR8383018.sra (13.42 GB)
- 5: SRR8383017.sra
- 4: SRR8383014.sra
- 3: SRR8383013.sra
- 2: zeo_mays_release-47_GTF.txt
- 1: zeo_mays_release-47_Genome.txt

After finishing the above four steps, four raw m⁶A sequencing reads named as **SRR8383013.sra, SRR8383013.sra, SRR8383013.sra, SRR8383013.sra**, respectively will be listed in your History Panel

3. Convert raw m⁶A sequencing reads from SRA to FASTQ format

Here, let's convert raw m⁶A sequencing reads from SRA to FASTQ format by the function **Sequencing Data Preprocessing** in **Data Preparation** module, please see the following four screenshots for details:

Step 1: process SRR8383013.sra

This step will cost ~12 minutes

Step 1: select SRR8383013.sra from History Panel

Step 2: select paired-end

Step 3: click here to run this function

The History panel on the right shows the processing status and results:

- 6: SRR8383018.sra (16.75 GB)
- 5: SRR8383017.sra
- 4: SRR8383014.sra
- 3: SRR8383013.sra
- 2: zeo_mays_release-47_GTF.txt
- 1: zeo_mays_release-47_Genome.txt

Step 2: process SRR8383014.sra

This step will cost ~26 minutes

Step 1: select SRR8383014.sra from History Panel

Step 2: select paired-end

Step 3: click here to run this function

Step 3: process SRR8383017.sra

This step will cost ~24 minutes

Step 1: select SRR8383017.sra from History Panel

Step 2: select paired-end

Step 3: click here to run this function

Step 4: process SRR8383018.sra

This step will cost ~23 minutes

Step 1: select SRR8383018.sra from History Panel

Step 2: select paired-end

Step 3: click here to run this function

For each SRA accession, two FASTQ format files (forward reads and reverse reads) will be generated as this experiment is paired-end sequencing.

4. Trim raw m⁶A sequencing reads

Before align m⁶A sequencing reads to genome, trim low-quality reads is necessary in all NGS (Next Generation Sequencing) analyses as which may cause incorrect mapping of reads to a reference genome, and even result in incorrect identification of RNA modifications. deepEA provided the function **Assess Reads Quality** in **Quality Control** module to filter raw reads to clean reads, please see the following screenshots for details:

Step 1: trim SRR8383013

This step will cost ~16 minutes

Step 1: select paired-end

Step 2: select forward read for SRR8383013

Step 3: select reverse read for SRR8383013

Step 4: set the minimum read length as 30

Step 5: click this button to run this function

Step 2: trim SRR8383014

This step will cost ~12 minutes

The screenshot shows the Galaxy web interface with the 'Assess Reads Quality' tool selected. The tool configuration is as follows:

- Single-end or paired-end reads?**: Paired-end
- Read file (R1)**: 9_SRR8383014.sra_1.fq
- Read file (R2)**: 10_SRR8383014.sra_2.fq
- Minimum read length**: 30
- Adapter sequences**: Auto detect
- Threads**: 1

Red annotations provide instructions:

- Step 1: select paired-end
- Step 2: select forward read for SRR8383014
- Step 3: select reverse read for SRR8383014
- Step 4: set the minimum read length as 30
- Step 5: click this button to run this function

The right side of the interface shows the 'History' panel with a list of datasets and their status.

Step 3: trim SRR8383017

This step will cost ~13 minutes

The screenshot shows the Galaxy web interface with the 'Assess Reads Quality' tool selected. The tool configuration is as follows:

- Single-end or paired-end reads?**: Paired-end
- Read file (R1)**: 9_SRR8383017.sra_1.fq
- Read file (R2)**: 10_SRR8383017.sra_2.fq
- Minimum read length**: 30
- Adapter sequences**: Auto detect
- Threads**: 1

Red annotations provide instructions:

- Step 1: select paired-end
- Step 2: select forward read for SRR8383014
- Step 3: select reverse read for SRR8383014
- Step 4: set the minimum read length as 30
- Step 5: click this button to run this function

The right side of the interface shows the 'History' panel with a list of datasets and their status.

Step 4: trim SRR8383018

This step will cost ~12 minutes

For each SRA accession, three files will be output, please see the following screenshot for detail:

5. Align clean m⁶A sequencing reads to *Zea mays* B73 reference genome with HISAT2

Here, we start to align clean reads to reference genome with HISAT2 provided in module **Identification of RNA Modifications**, see the following screenshots for details:

Step 1: align SRR8383013

This step will cost ~48 minutes

Step 1: select *Zea mays* reference genome

Step 2: select paired-end

Step 3: select clean forward read for SRR8383013

Step 4: select clean reverse read for SRR8383013

Step 5: select not to extract uniquely mapped reads

Step 6: click this button to run this function

The Galaxy interface shows the HISAT2 tool configuration. The 'Reference genome sequence' input field is set to '1: zea_mays_release-47_Genome.txt'. The 'Single or paired library' dropdown is set to 'Paired-end'. The 'Read file (R1)' input field contains '16: SRR8383013.sra_1.fq_clean_reads_R1.fastq'. The 'Read file (R2)' input field contains '17: SRR8383013.sra_2.fq_clean_reads_R2.fastq'. The 'Algorithm options' section has 'Extract uniquely mapped reads?' set to 'No'. The 'Execute' button is highlighted with a red box.

Step 2: align SRR8383014

This step will cost ~40 minutes

Step 1: select *Zea mays* reference genome

Step 2: select paired-end

Step 3: select clean forward read for SRR8383014

Step 4: select clean reverse read for SRR8383014

Step 5: select not to extract uniquely mapped reads

Step 6: click this button to run this function

The Galaxy interface shows the HISAT2 tool configuration. The 'Reference genome sequence' input field is set to '1: zea_mays_release-47_Genome.txt'. The 'Single or paired library' dropdown is set to 'Paired-end'. The 'Read file (R1)' input field contains '19: SRR8383014.sra_1.fq_clean_reads_R1.fastq'. The 'Read file (R2)' input field contains '20: SRR8383014.sra_2.fq_clean_reads_R2.fastq'. The 'Algorithm options' section has 'Extract uniquely mapped reads?' set to 'No'. The 'Execute' button is highlighted with a red box.

Step 3: align SRR8383017

This step will cost ~44 minutes

Step 1: select Zea mays reference genome

Step 2: select paired-end

Step 3: select clean forward read for SRR8383017

Step 4: select clean reverse read for SRR8383017

Step 5: select not to extract uniquely mapped reads

Step 6: click this button to run this function

Step 4: align SRR8383018

This step will cost ~42 minutes

Step 1: select Zea mays reference genome

Step 2: select paired-end

Step 3: select clean forward read for SRR8383018

Step 4: select clean reverse read for SRR8383018

Step 5: select not to extract uniquely mapped reads

Step 6: click this button to run this function

Then for each SRA accession, reads-genome alignments in BAM format and alignment summary will be generated, see the following screenshot:

The screenshot shows the Galaxy web interface with the HISAT2 tool selected. The main panel displays the tool configuration for aligning reads from SRR8383018.sra and SRR8383017.sra against the zea_mays_release-47_Genome.txt reference genome. The 'Paired-end' option is chosen, and the 'Basic' algorithm is selected. The 'Extract uniquely mapped reads?' checkbox is checked. A red box highlights the 'Execute' button.

Reads-genome alignments of SRR8383013 in BAM format

Alignment summary

Click here to view the alignment summary

The right panel shows the generated history, which includes multiple entries for HISAT2 alignments and their corresponding BAM files and alignment summaries. A red box highlights the first entry for SRR8383013.sra.

6. Call m⁶A enriched peaks with macs2

After finishing aligning reads to genome, let's start to call m⁶A enriched peaks with macs2, the following screenshots show details about parameter settings:

Step 1: call m⁶A peaks for biological replicate 1

This step will cost ~8 minutes

The screenshot shows the Macs2 tool configuration for peak calling. The 'Input the BAM file in sample' field contains 'hisat2_alignment.bam (on SRR8383013.sra_1.fq_clean_reads_R1.fastq and SRR8383013.sra_2.fq_clean_reads_R2.fastq)'. The 'Input the BAM file in input sample' field contains 'hisat2_alignment.bam (on SRR8383017.sra_1.fq_clean_reads_R1.fastq and SRR8383017.sra_2.fq_clean_reads_R2.fastq)'. The 'Peak calling methods' section has 'MACS2' selected. A red box highlights the 'MACS2' selection. The 'Effective genome size' field is set to '41727442'. A red box highlights this value. The 'Build Model' section shows 'Set lower mfold bound' at '5' and 'Set upper mfold bound' at '50'. A red box highlights the '5' value. The 'Band width for picking regions to compute fragment size' field is set to '300'. A red box highlights this value. The 'Peak detection based on' dropdown is set to 'q-value'. The 'Default uses: q-value' dropdown is also set to 'q-value'. The 'Minimum FDR (q-value) cutoff for peak detection' field is set to '0.01'. A red box highlights this value. The 'Execute' button is at the bottom. A red box highlights it.

Step 1: input IP and Input alignment respectively for biological replicate 1

Step 2: select macs2 to call peaks

Step 3: set effective genome size

Step 4: set significant threshold

Step 5: click here to run this function

The right panel shows the generated history, which includes multiple entries for Macs2 alignments and their corresponding BAM files and alignment summaries. A red box highlights the first entry for SRR8383013.sra.

This step will generate ~16,400 peaks for biological replicate 1

Step 2: call m⁶A peaks for biological replicate 2

This step will cost ~7 minutes

This step will generate ~16,100 peaks for biological replicate 2

Step 3: obtain consistent peaks between two biological replicates

After finishing peak calling for two replicates, deepEA also provided a function **Merge two biological replicates** to obtain consistent peaks between two biological replicates, see the following screenshots for details:

This step will cost ~2 seconds

Then the consistent peaks named as **intersect.bed** (about ~14,000 consistent peaks) will be shown in your History Panel.

7. Perform functional annotation for m⁶A

m⁶A distribution

The distribution of m⁶A in the genome and transcriptome can be visualized by the function **RNA Modification Distribution** in **Functional Annotation** module, see the following screenshot for detail:

This step will cost ~5 minutes

Step 1: select consistent peaks from history
Step 2: select genome annotation
Step 3: click here to run this function

Then an interactively HTML will be generated, please see [here](#) to preview this results

Link m⁶A modifications with genes

This step will cost ~2 minutes

Step 1: consistent m⁶A peaks between two biological replicates
Step 2: input genome annotation
Step 3: click here to run this function

This output for this function are shown in the following the screenshot

The screenshot shows the Galaxy web interface. On the left, the 'Tools' sidebar lists various analysis categories like PRE-ANALYSIS, CORE ANALYSIS, ADVANCED ANALYSIS, and USEFUL TOOLS. In the center, a table displays genomic data with columns: Chrom, Start, End, Name, Score, Strand, ThickStart, ThickEnd, ItemKB, BlockCount, BlockSizes, and BlockStarts. A red arrow points from the top of this table to the 'History' panel on the right. The 'History' panel shows a list of workflow steps, with the last two steps highlighted in green: '40: RNA_modifications_with_strand.bed' and '39: RNA_modifications_genes.txt'. Below these are other steps: '38: RNA_modifications_distribution.html', '37: intersect.bed', '36: MACS2_peaks.bed', '35: MACS2_peaks.bed', '34: hisat2_alignment.bam (on SR8R832301 & sra_1.fq_clean_reads_R1.fastq and SR8R833018.sra_2.fq_clean_reads_R2.fastq)', '33: hisat2_alignment.summary.txt (on SR8R832301 & sra_1.fq_clean_reads_R1.fastq and SR8R833018.sra_2.fq_clean_reads_R2.fastq)', '32: hisat2_alignment.bam (on SR8R832301 & sra_1.fq_clean_reads_R1.fastq and SR8R833017.sra_2.fq_clean_reads_R2.fastq)', and '31: hisat2_alignment.summary.txt (on SR8R832301 & sra_1.fq_clean_reads_R1.fastq and SR8R833017.sra_2.fq_clean_reads_R2.fastq)'. The top of the history panel indicates 'Using 591.2 GB'.

De-novo motif discovery

This following screenshot shows how to use homer to perform *de-novo* motif discovery

This step will cost ~6 minutes

The screenshot shows the 'Motif Analysis (Galaxy Version 17.09)' tool configuration page. It includes fields for 'RNA modifications (peak regions or single nucleotide resolution) in BED format' (containing '40: RNA_modifications_with_strand.bed'), 'Reference genome sequences in FASTA format' (containing '1: zea_mays_release-47_Genome.txt'), 'Select a method' (radio buttons for HOMER, MEME-ChIP, and DREME), 'Options Configuration' (radio buttons for Basic and Advanced, and a 'motif length' input field set to '5,6,7,8'), and 'Execute' (a blue button). Red annotations with arrows point to each of these fields with the labels: Step 1: input peak regions with strand information, Step 2: input reference genome sequences, Step 3: set the motif length, and Step 4: click here to run this function. To the right of the form is a 'History' panel showing a list of workflow steps, many of which are highlighted in green, indicating they have been completed. The top of the history panel indicates 'Using 591.2 GB'.

Then an HTML document will be generated, please click [here](#) to preview this results

GO functional enrichment analysis

deepEA provided the function **Functional Enrichment Analysis** to perform GO enrichment analysis, see the following screenshot for detail:

This step will cost ~1 minute

The screenshot shows the Galaxy Functional Enrichment Analysis tool interface. Step 1: A red box highlights the 'Latin species name' input field, which contains 'Latin species name'. Step 2: A red box highlights the 'RNA modifications gene list' input field, which contains '39 RNA_modifications_genes.txt'. Step 3: A red box highlights the 'Execute' button at the bottom left of the form.

Step 1: input latin species name

Step 2: select RNA modifications gene list (generated by function Link RNA Modifications to Genes)

Step 3: click here to run this function

Then a figure in PDF format and a TAB separated matrix will be generated, please click [here](#) for details.

7. Multi-omics integrative analysis

To run this module, you have to download test data provided by deepEA, and then upload the data in directory `test_data/Multi-omics Integrative Analysis/` to deepEA server. If you are not sure how to upload local data into deepEA server, please see [here](#) for details. Then you can run the function **Integrative Analysis of Three Omics Data Sets** in **Multi-omics Integrative Analysis** module as the following screenshot shows:

This step will cost ~6 seconds

The screenshot shows the Galaxy Integrative Analysis of Three Omics Data Sets tool interface. Step 1: A red box highlights the 'quantification_matrix' input field, which contains '53: quantification_matrix_three_omics.txt'. Step 2: A red box highlights the 'Duplicated gene pairs (homoeologs)' input field, which contains '54: Duplicated-gene-pairs.txt'. Step 3: A red box highlights the 'Execute' button at the bottom left of the form.

Step 1: input the quantification matrix

Step 2: input duplicated gene pairs

Step 3: click here to run this function

Then an interactively HTML document will be output, click [here](#) to preview this results.

8. Build an m⁶A predictor based on machine learning

The following four steps shows us how to use **Prediction Analysis Based on Machine Learning** module to build an m⁶A predictor and perform cross-validation experiments.

Step 1: generate positive and negative samples for m⁶A predictor construction

The function **Sample Generation** can be used to generate positive and negative samples.

This step will cost ~5 hours, please wait patiently

Step 1: input peak regions with strand information
Step 2: input reference genome sequences
Step 3: input genome annotation
Step 4: select Yes to generate negative samples
Step 5: click here to run this function

Step 2: encoding positive samples by function Feature Encoding

This step will cost ~13 minutes

Step 1: select positive samples from History Panel
Step 2: select Zea mays reference genome sequence
Step 3: select feature encoding methods
Step 4: input Zea mays genome annotation
Step 5: select nucleic acid composition related features
Step 6: select autocorrelation-based features (GAC is not included as it runs particularly slow)
Step 7: select all features here
Step 8: click here to run this function

Step 3: encoding negative samples by function Feature Encoding

This step will cost ~13 minutes

Step 1: select negative samples from History Panel

Step 2: select Zea mays reference genome sequence

Step 3: select feature encoding methods

Step 4: input Zea mays genome annotation

Step 5: select nucleic acid composition related features

Step 6: select autocorrelation-based features (GACG is not included as it runs particularly slow)

Step 7: select all features here

Step 8: click here to run this function

Step 4: m⁶A predictor construction and evaluation

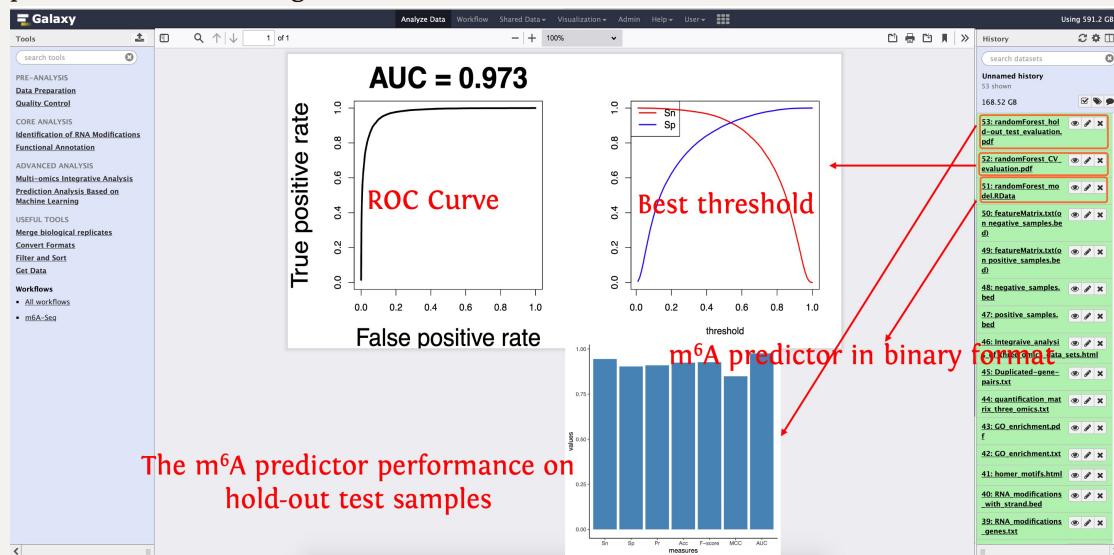
Step 1: select feature matrix of positive samples

Step 2: select feature matrix of negative samples

Step 3: set the number of threads as 5

Step 4: click here to run this function

After finishing predictor construction and evaluation, three documents will be output, please see the following screenshot:

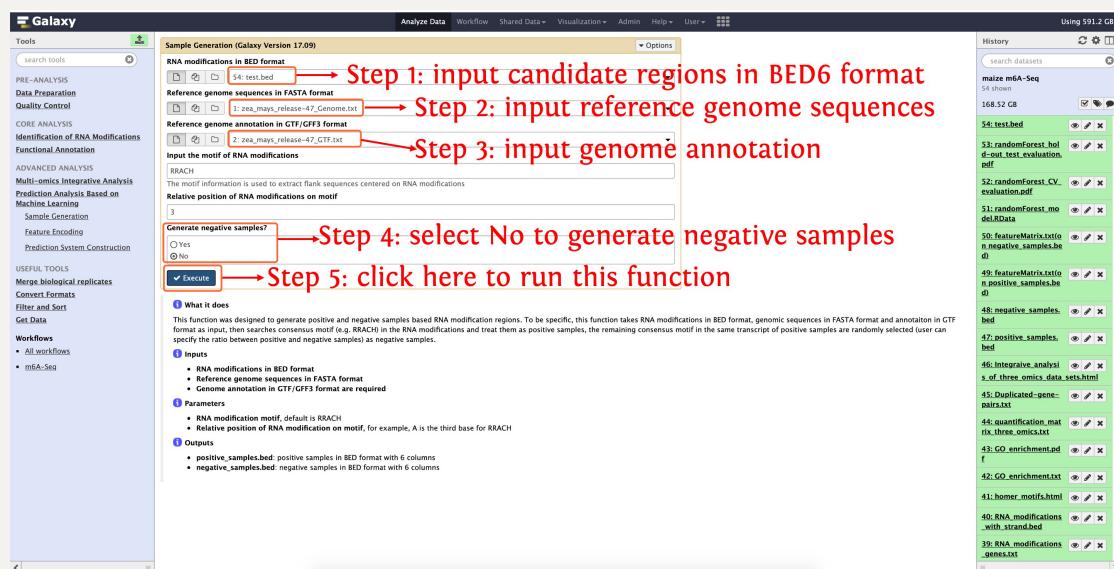


9. Predict candidate m⁶A using the m⁶A predictor

After building an m⁶A predictor, we can use the predictor to prioritize candidate m⁶ modifications, to do this, you have to prepare a file in BED format, see the following table:

| CHR | START | END | GENEID | NA | STRAND |
|-----|-------|-------|----------------|----|--------|
| 1 | 49625 | 49751 | Zm00001d027230 | . | + |
| 1 | 50925 | 51026 | Zm00001d027231 | . | - |
| 1 | 92303 | 92526 | Zm00001d027232 | . | - |

Then upload this file into deepEA server and run the function **Sample Generation** in **Prediction Analysis Based on Machine Learning** module as the following screenshot shows:



Then encoding the candidate samples using the same feature encoding methods as m⁶A predictor construction:

The screenshot shows the Galaxy Feature Encoding tool interface. A series of red arrows and annotations guide the user through the configuration process:

- Step 1:** Select samples generated by last function (highlighted in red).
- Step 2:** Select Zea mays reference genome sequence (highlighted in red).
- Step 3:** Select feature encoding methods (highlighted in red).
- Step 4:** Input Zea mays genome annotation (highlighted in red).
- Step 5:** Select nucleic acid composition related features (highlighted in red).
- Step 6:** Select autocorrelation-based features (GAC is not included as it runs particularly slow) (highlighted in red).
- Step 7:** Select all features here (highlighted in red).
- Step 8:** Click here to run this function (highlighted in red).

The History panel on the right shows a list of completed steps, including "maize m6A-Seq", "randomForest.hol", "randomForest.CV", "randomForest.m", "featureMatrix.txt", "negative.samples.bed", "positive.samples.bed", "integrate.analysis", "three_omics.data_sets.html", "Duplicated-gene-pairs.txt", "quantification_matrix.three_omics.txt", "GO_enrichment.txt", "homologous.txt", and "RNA_modification_with_strand.bed".

After encoding the candidate samples, we can predict the candidate m⁶A:

The screenshot shows the Galaxy Prediction System Construction tool interface. A series of red arrows and annotations guide the user through the configuration process:

- Step 1:** Select to predict (highlighted in red).
- Step 2:** Input feature matrix of candidate samples (highlighted in red).
- Step 3:** Input the trained model (highlighted in red).
- Step 4:** Select Random Forest (highlighted in red).
- Step 5:** Click here to run this function (highlighted in red).

The History panel on the right shows a list of completed steps, including "maize m6A-Seq", "positive.samples.bed", "test.bed", "randomForest.hol", "randomForest.CV", "randomForest.m", "featureMatrix.txt", "negative.samples.bed", "positive.samples.bed", "integrate.analysis", "three_omics.data_sets.html", "Duplicated-gene-pairs.txt", "quantification_matrix.three_omics.txt", "GO_enrichment.txt", "homologous.txt", and "RNA_modification_with_strand.bed".