HEMU

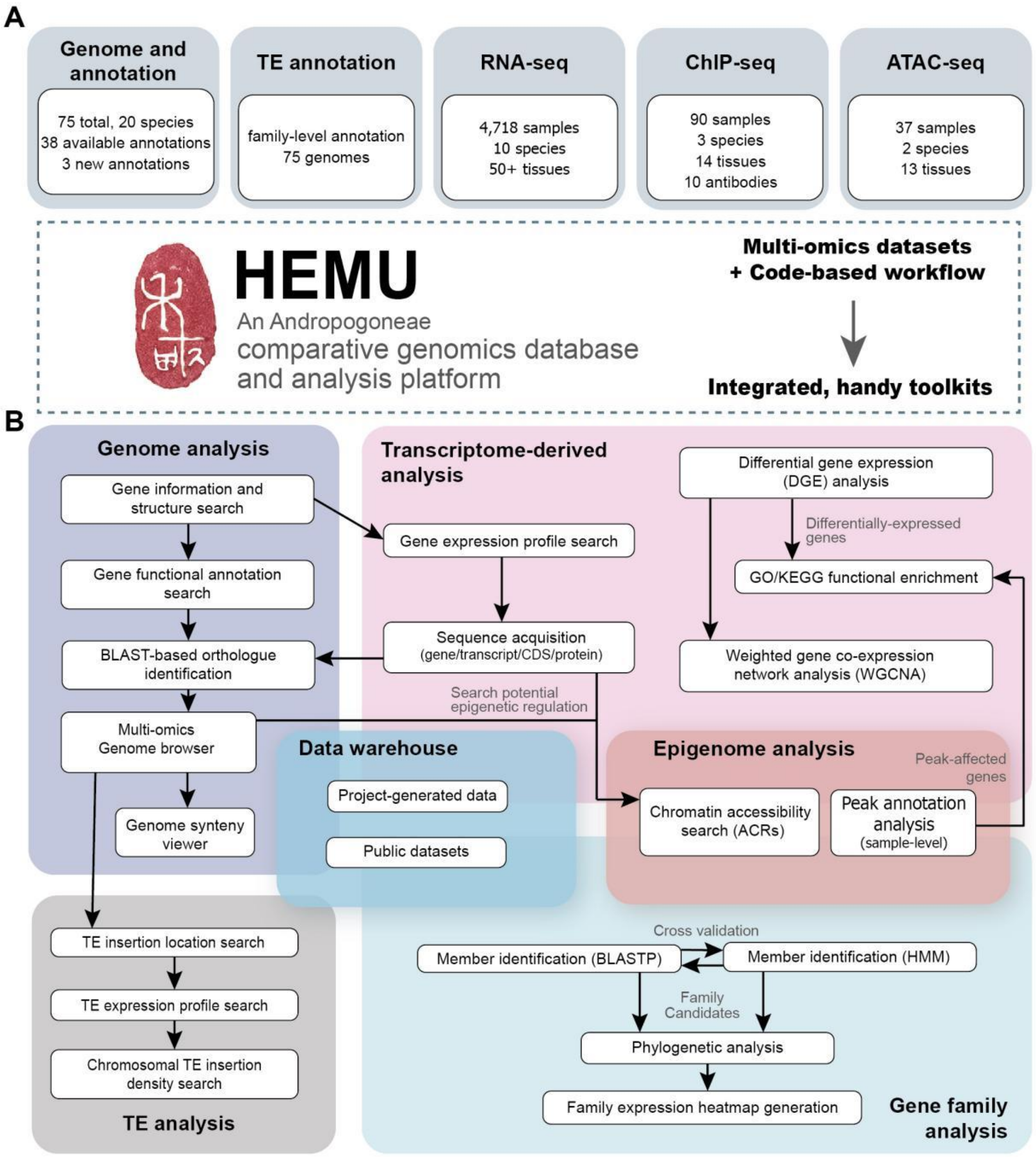
database

Contents

目录

Chapter 1

1 Introduction


A
Genome and annotation
75 total, 20 species
38 available annotations
3 new annotations
TE annotation
family-level annotation
75 genomes
RNA-seq
4,718 samples
10 species
50+ tissues
ChIP-seq
90 samples
3 species
14 tissues
10 antibodies
ATAC-seq
37 samples
2 species
13 tissues
HEMU
An Andropogoneae
comparative genomics database
and analysis platform
Multi-omics datasets
+ Code-based workflow
Integrated, handy toolkits
B
Genome analysis
Gene information and structure search
Gene functional annotation search
BLAST-based orthologue identification
Multi-omics Genome browser
Genome synteny viewer
Transcriptome-derived analysis
Gene expression profile search
Sequence acquisition (gene/transcript/CDS/protein)
Search potential epigenetic regulation
Differential gene expression (DGE) analysis
Differentially-expressed genes
GO/KEGG functional enrichment
Weighted gene co-expression network analysis (WGCNA)
Data warehouse
Project-generated data
Public datasets
Epigenome analysis
Peak-affected genes
Chromatin accessibility search (ACRs)
Peak annotation analysis (sample-level)
TE insertion location search
TE expression profile search
Chromosomal TE insertion density search
TE analysis
Cross validation
Member identification (BLASTP)
Member identification (HMM)
Family Candidates
Phylogenetic analysis
Family expression heatmap generation
Gene family analysis

1.1 What is HEMUdb?

After reviewing the existing literature on Andropogoneae, it was evident that a comprehensive database for the primary species within the Andropogoneae group was lacking. Consequently, we developed HEMUdb, a user-friendly and comprehensive database tailored for researchers focusing on Andropogoneae.

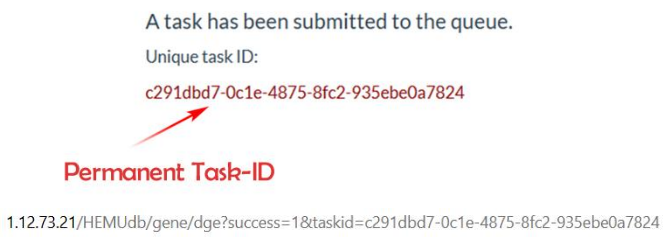
1.2 What do we provide?

We have developed a user-friendly graphical interface that allows users to effortlessly explore genomic data across various representative Andropogoneae species with just a click. Leveraging a total of 4287 RNA-seq datasets from the public database, we applied a widely recognized transcriptome analysis process to construct 73 significant genomes of the Andropogoneae tribe.

We have designed four distinct search toolkits for users to utilize. Our server infrastructure is built on Django, MySQL, and Shiny technologies. You can conveniently access various types of visual analysis results (further details provided below). Should you have any inquiries or suggestions regarding our HEMUdb, feel free to contact us via email at **zhuyzh37@mail2.sysu.edu.cn** (Edward Zhu).

1.3 A useful function — Task ID

For enhanced user convenience, a distinctive Task ID will be assigned during the execution of the process. This Task ID will be visible in the URL and will remain effective indefinitely. Therefore, you have the option to save the Task ID and retrieve its associated results whenever needed.



A task has been submitted to the queue.

Unique task ID:

c291dbd7-0c1e-4875-8fc2-935ebe0a7824

Permanent Task-ID

1.12.73.21/HEMUdb/gene/dge?success=1&taskid=c291dbd7-0c1e-4875-8fc2-935ebe0a7824

Figure 1.3.1 an example of Task-ID and its place on URL

Chapter 2

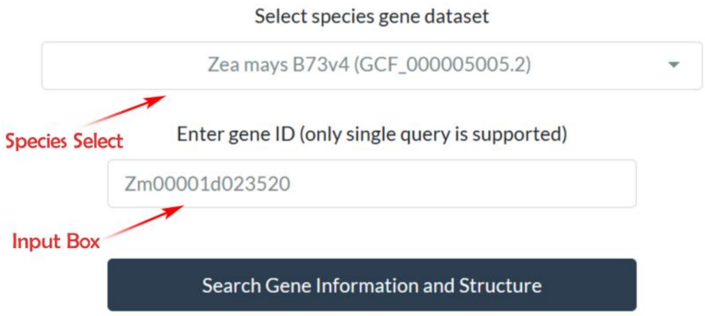
2 Toolkit I: Genome

2.1 Gene Information & Structure Search

We have developed a query page using the Shiny framework, allowing you to explore the gene structure of representative Andropogoneae species. To access this information, follow these straightforward steps:

- **Enter the gene ID you're interested in.**
- **Choose the relevant species.**
- **Click the Search button and wait for the results to load.**

This will provide you with details about the major transcript, CDS (Coding Sequences), UTR (Untranslated Regions), and other structural information associated with the specified gene.


Select species gene dataset

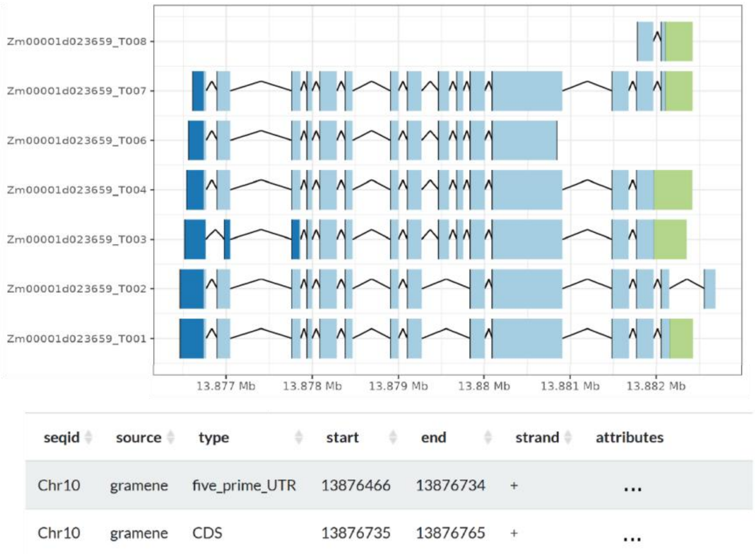
Zea mays B73v4 (GCF_000005005.2)

Enter gene ID (only single query is supported)

Zm00001d023520

Figure 2.1.1 interface of Gene Structure Visualization Search

You can track the query's progress and results loading by observing the progress bar located at the bottom right corner of the page. Once the loading is finished, the generated results will be displayed on the same page.



seqid	source	type	start	end	strand	attributes
Chr10	gramene	five_prime_UTR	13876466	13876734	+	...
Chr10	gramene	CDS	13876735	13876765	+	...

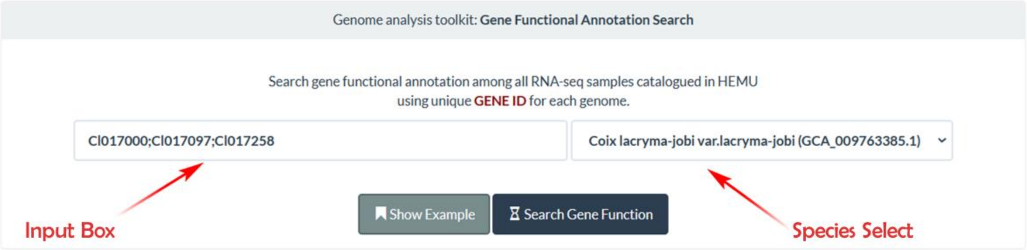
Figure 2.1.2 results of Gene Structure Visualization and details of each gene element

The initial figure in the results will provide a graphical representation of the primary transcript details for the gene, with distinct gene structures represented by various colors. The number of displayed transcripts can be modified using the option located in the lower left corner. Subsequently, a table will present comprehensive information about each gene element. This information encompasses attributes such as source, start/end positions, and more.

2.2 Gene Functional Annotation Search

We have analyzed the gene functional annotation data for selected Andropogoneae species. As a result, we've created a user-friendly interface to search through the annotation outcomes. You can explore gene functional annotations across RNA-seq samples that are part of the HEMU catalog. To do this, you'll need to use the specific Gene ID for each genome. Here's how you can proceed:

- **Enter the Gene ID(s) you are interested in (separated by semicolons).**
- **Choose the relevant species from the provided options.**
- **Click on the "Search Gene Function" button and wait for the results to appear.**



Genome analysis toolkit: **Gene Functional Annotation Search**

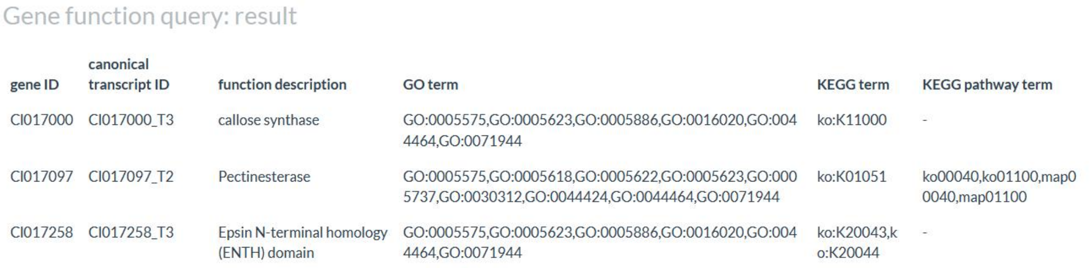
Search gene functional annotation among all RNA-seq samples catalogued in HEMU using unique **GENE ID** for each genome.

CI017000;CI017097;CI017258

Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)

Figure 2.2.1 interface of Functional Analysis

After a brief waiting period of a few seconds, the results will be presented within the same interface. A table will display comprehensive information, encompassing the primary transcript ID, a concise functional description, as well as GO/KEGG term and KEGG pathway term details.



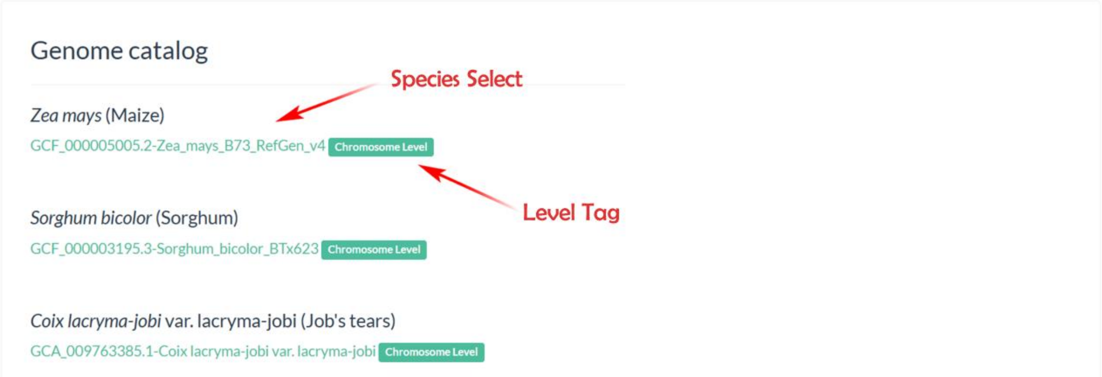
Gene function query: result

gene ID	canonical transcript ID	function description	GO term	KEGG term	KEGG pathway term
CI017000	CI017000_T3	callose synthase	GO:0005575,GO:0005623,GO:0005886,GO:0016020,GO:0044464,GO:0071944	ko:K11000	-
CI017097	CI017097_T2	Pectinesterase	GO:0005575,GO:0005618,GO:0005622,GO:0005623,GO:0005737,GO:0030312,GO:0044424,GO:0044464,GO:0071944	ko:K01051	ko00040,ko01100,map00040,map01100
CI017258	CI017258_T3	Epsin N-terminal homology (ENTH) domain	GO:0005575,GO:0005623,GO:0005886,GO:0016020,GO:0044464,GO:0071944	ko:K20043,ko:K20044	-

Figure 2.2.2 results of Gene Function Query

2.3 Multi-omics Genome Browser

Multi-omics analysis is a systems biology approach that comprehensively explores molecular-level information within organisms by integrating various high-throughput technologies. This holistic method enables researchers to gain a more comprehensive and in-depth understanding of the interactions and regulatory relationships among various molecules within a biological system. This page is designed to showcase visualized genomic data for individual species that have been processed. A total of 9 species are available, and custom parameters can be applied to visualize Multi-omics genome data in a linear view. The genome assemblies of six species have reached the chromosome level. Click on the green link below each species name to view the corresponding visualization. For a more immersive experience, click the "Go fullscreen" button at the top of the page.



Genome catalog

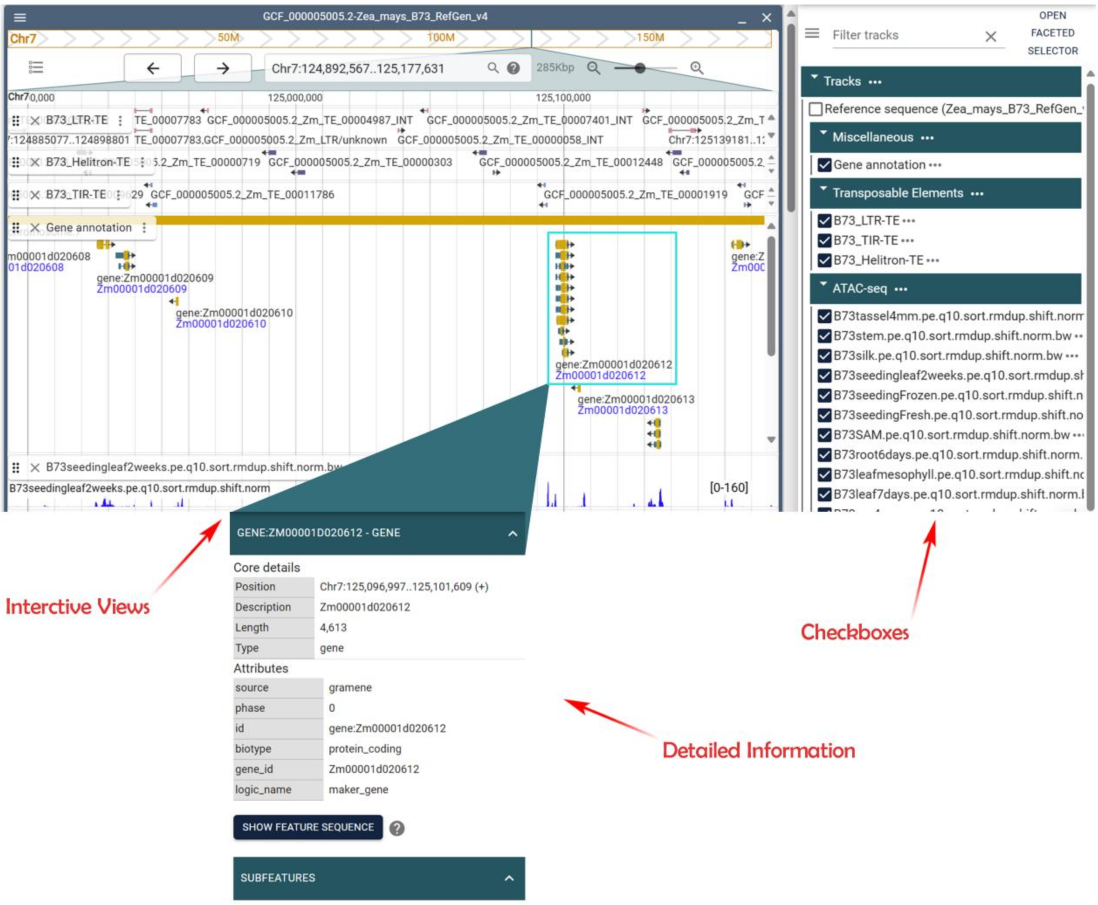
Zea mays (Maize)
GCF_000005005.2-Zea_mays_B73_RefGen_v4 Chromosome Level

Sorghum bicolor (Sorghum)
GCF_000003195.3-Sorghum_bicolor_BTx623 Chromosome Level

Coix lacryma-jobi var. lacryma-jobi (Job's tears)
GCA_009763385.1-Coix lacryma-jobi var. lacryma-jobi Chromosome Level

Figure 2.3.1 interface of Multi-omics Genome Browser

Taking the example of Chromosome 7 of *Zea mays*, genomic data, including gene annotation, transposable elements, and ATAC-seq, is presented on an interactive page. You can select specific genomic information of interest from the checkboxes on the right. The most recently chosen content will be dynamically generated at the bottom of the page. Clicking on genes within the page provides more detailed information. Additionally, you can use other action buttons at the top of the page to adjust the displayed windows.


GCF_000005005.2-Zea_mays_B73_RefGen_v4
Chr7
50M
100M
150M
Chr7:124,892,567..125,177,631
285Kbp
125,000,000
125,100,000
B73_LTR-TE
B73_Helitron-TE
B73_TIR-TE
Gene annotation
gene:Zm00001d020609
gene:Zm00001d020610
gene:Zm00001d020612
gene:Zm00001d020613
B73seedingleaf2weeks.pe.q10.sort.rmdup.shift.norm.bw
[0-160]
Filter tracks
OPEN
FACETED
SELECTOR
Tracks
□ Reference sequence (Zea_mays_B73_RefGen_
Miscellaneous
☑ Gene annotation
Transposable Elements
☑ B73_LTR-TE
☑ B73_TIR-TE
☑ B73_Helitron-TE
ATAC-seq
☑ B73tassel4mm.pe.q10.sort.rmdup.shift.norm
☑ B73stem.pe.q10.sort.rmdup.shift.norm.bw
☑ B73silk.pe.q10.sort.rmdup.shift.norm.bw
☑ B73seedingleaf2weeks.pe.q10.sort.rmdup.sh
☑ B73seedingFrozen.pe.q10.sort.rmdup.shift.n
☑ B73seedingFresh.pe.q10.sort.rmdup.shift.no
☑ B73SAM.pe.q10.sort.rmdup.shift.norm.bw
☑ B73root6days.pe.q10.sort.rmdup.shift.norm.
☑ B73leafmesophyll.pe.q10.sort.rmdup.shift.nc
☑ B73leaf7days.pe.q10.sort.rmdup.shift.norm.l
SHOW FEATURE SEQUENCE
SUBFEATURES

Interctive Views

Checkboxes

Detailed Information

GENE:ZM00001D020612 - GENE

Core details

Position	Chr7:125,096,997..125,101,609 (+)
Description	Zm00001d020612
Length	4,613
Type	gene

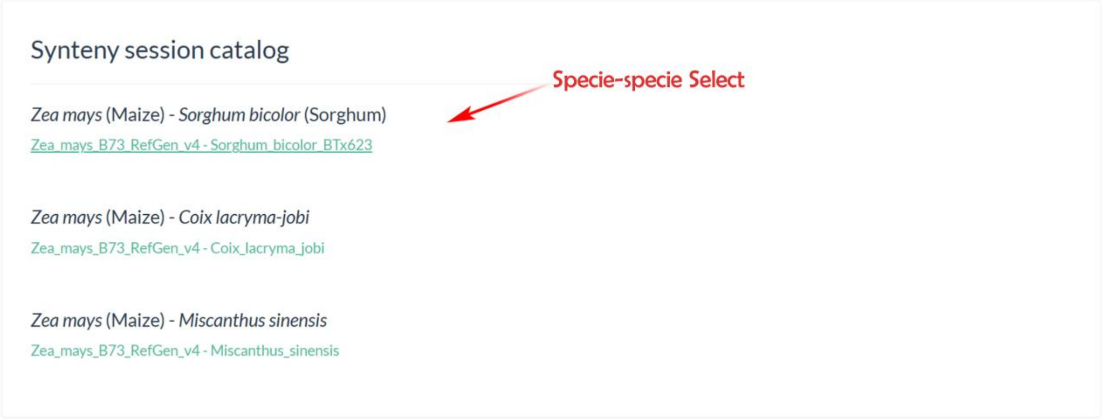
Attributes

source	gramene
phase	0
id	gene:Zm00001d020612
biotype	protein_coding
gene_id	Zm00001d020612
logic_name	maker_gene

Figure 2.3.2 interactive interface of multi-omics genomic information

2.4 Genome Synteny Viewer

Genome synteny refers to the phenomenon of the conservation of the order and orientation of genes or other genetic elements across the genomes of different species. Synteny provides insights into the evolutionary relationships between species and can reveal the conservation of genetic content and organization. Currently, we have completed synteny analyses among three pairs of species and have visualized the results for user convenience. Click on the green link below each species to view the synteny results for that specific pair of species.



Synteny session catalog

Specie-specie Select

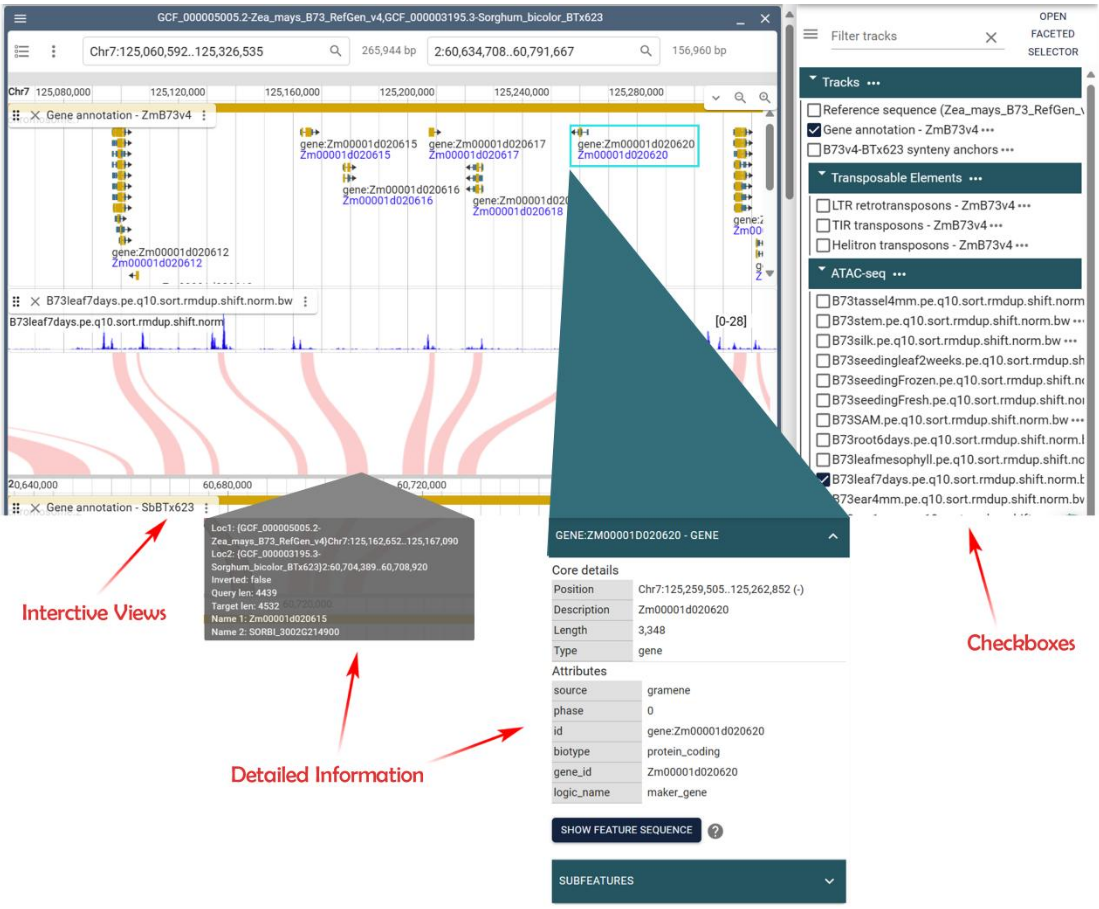
Zea mays (Maize) - *Sorghum bicolor* (Sorghum)
Zea_mays_B73_RefGen_v4 - Sorghum_bicolor_BTx623

Zea mays (Maize) - *Coix lacryma-jobi*
Zea_mays_B73_RefGen_v4 - Coix_lacryma_jobi

Zea mays (Maize) - *Miscanthus sinensis*
Zea_mays_B73_RefGen_v4 - Miscanthus_sinensis

Figure 2.4.1 interface of Genome Synteny Viewer

Taking the example of chromosome 7 in *Zea mays* and chromosome 2 in *Sorghum bicolor*, in addition to various omics data such as genome annotation, transposable elements, ATAC-seq, etc., a synteny plot showing the corresponding positions of similar sequences between the chromosomes of the two species will be generated. Additionally, a set of Synteny anchors will also be generated by default. The framework used by viewers is the same as mentioned earlier, and relevant operational functions can be referred to in the methods described in the previous section on the **Multi-omics Genome Browser**.



Tracks

□ Reference sequence (Zea_mays_B73_RefGen_v

☑ Gene annotation - ZmB73v4

□ B73v4-BTx623 synteny anchors

Transposable Elements

□ LTR retrotransposons - ZmB73v4

□ TIR transposons - ZmB73v4

□ Helitron transposons - ZmB73v4

ATAC-seq

□ B73tassel4mm.pe.q10.sort.rmdup.shift.norm

□ B73stem.pe.q10.sort.rmdup.shift.norm.bw

□ B73silk.pe.q10.sort.rmdup.shift.norm.bw

□ B73seedingleaf2weeks.pe.q10.sort.rmdup.sh

□ B73seedingFrozen.pe.q10.sort.rmdup.shift.n

□ B73seedingFresh.pe.q10.sort.rmdup.shift.no

□ B73SAM.pe.q10.sort.rmdup.shift.norm.bw

□ B73root6days.pe.q10.sort.rmdup.shift.norm.I

□ B73leafmesophyll.pe.q10.sort.rmdup.shift.nc

☑ B73leaf7days.pe.q10.sort.rmdup.shift.norm.b

B73ear4mm.pe.q10.sort.rmdup.shift.norm.bv

```
Loc1: {GCF_000005005.2-
Zea_mays_B73_RefGen_v4}Chr7:125,162,652..125,167,090
Loc2: {GCF_000003195.3-
Sorghum_bicolor_BTx623}2:60,704,389..60,708,920
Inverted: false
Query len: 4439
Target len: 4532
Name 1: Zm00001d020615
Name 2: SORBI_3002G214900
```

GENE:ZM00001D020620 - GENE

Core details

Position	Chr7:125,259,505..125,262,852 (-)
Description	Zm00001d020620
Length	3,348
Type	gene

Attributes

source	gramene
phase	0
id	gene:Zm00001d020620
biotype	protein_coding
gene_id	Zm00001d020620
logic_name	maker_gene

Interctive Views

Detailed Information

Checkboxes

Figure 2.4.2 interactive interface of specie-specie genomic data information

2.5 The BLAST Server

We have incorporated part of the BLAST search engine framework from Sequance Server (sequenceserver.com) and supplemented it with nucleotide and protein databases assembled for our species. You can either directly access the page to input sequences of interest for analysis or, alternatively, click on "Send sequence to BLAST module" from almost any page that generates sequences to be redirected to this analysis page. By selecting the relevant database information, you can obtain alignment results.

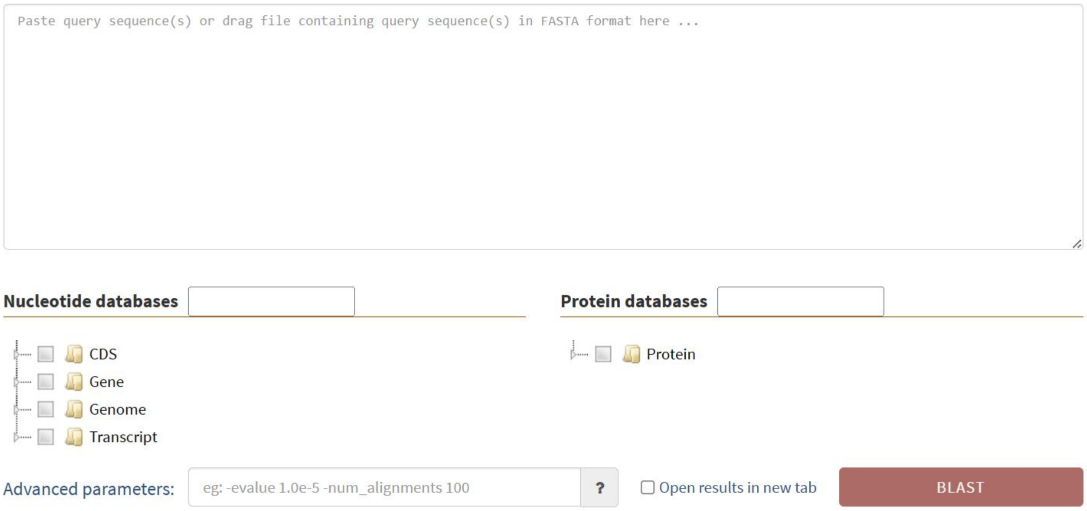

Paste query sequence(s) or drag file containing query sequence(s) in FASTA format here ...
Nucleotide databases
Protein databases
CDS
Gene
Genome
Transcript
Protein
Advanced parameters:
eg: -evalue 1.0e-5 -num_alignments 100
?
Open results in new tab
BLAST

Figure 2.5.1 interface of BLAST module

Chapter 3

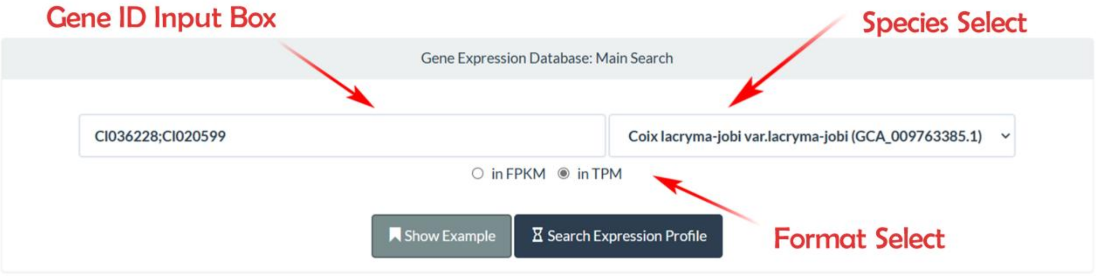
3 Toolkit II: Transcriptome

3.1 Gene Expression Profiles

We have introduced the **Online Query Module** for gene expression in the Andropogoneae transcriptome database. The search interface is shown in Figure 3.1.1. To obtain gene expression profiles, follow these steps:

- **Enter the Gene IDs you are interested in into the left input box (use ";" to separate multiple samples).**
- **Select the desired Andropogoneae family from the options on the right.**
- **Choose your preferred expression format (FPKM and TPM are both supported).**

If you are using this function for the first time, you can click on the "Show Example" button. This will automatically fill in a prepared Gene ID in the left box.


Gene ID Input Box
Species Select
Format Select
Gene Expression Database: Main Search
Show Example
Search Expression Profile

CI036228;CI020599

Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)

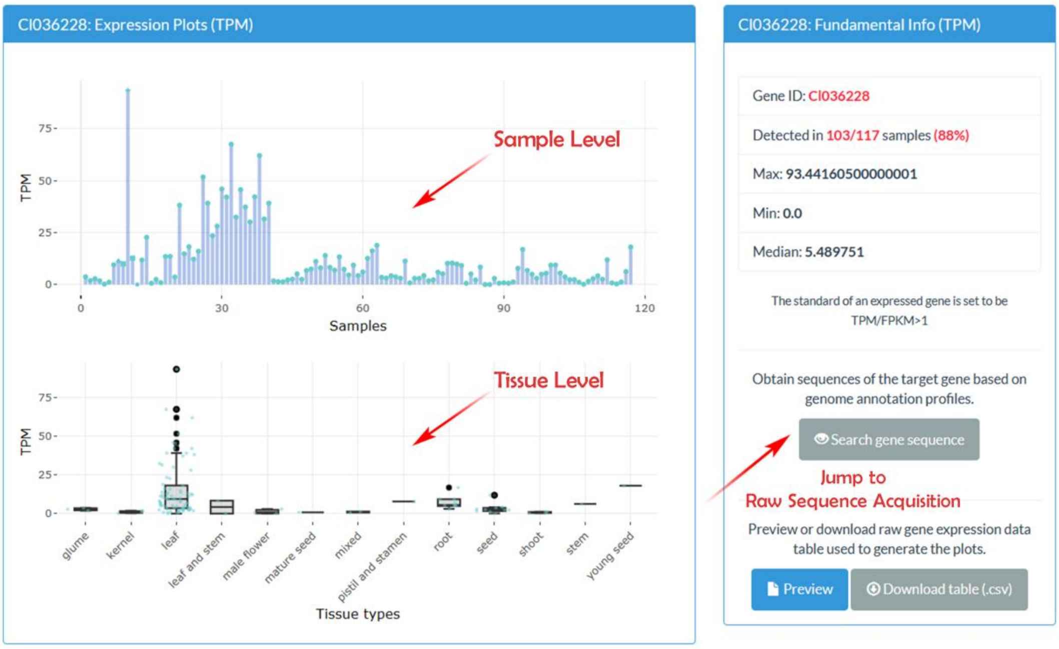
○ in FPKM ● in TPM

Figure 3.1.1 interface of Gene Expression Profiles

After a brief wait of a few seconds, you will be able to access the results, presented through two panels showcasing gene expression profiles for each Gene ID.

Expression Plots We offer two interactive expression plots to visualize the profile outcomes, encompassing both sample-level and tissue-level data. In the sample-level plots, by hovering your cursor over any sample data point, you can view its corresponding ID, along with the associated TPM/FPKM value and sample ID. Likewise, within the tissue-level plots, you can employ the same approach to retrieve information about the tissue type and its corresponding TPM/FPKM value for any given sample data point.

Fundamental Info This section provides essential information about the queried gene, including frequency, maximum, minimum, and median expression levels across samples. This furnishes you with a fundamental understanding of the Gene ID you're interested in. By clicking the "Search gene sequence" button, you will be directed to the Raw Sequence Acquisition page, enabling you to access sequences for the target gene based on genome annotation profiles. Additionally, you have the option to preview or download the raw gene expression data table, which was used to generate the plots. This table supplies Sample IDs, TPM and FPKM values, as well as tissue types.



Gene ID: CI036228

Detected in 103/117 samples (88%)

Max: 93.44160500000001

Min: 0.0

Median: 5.489751

The standard of an expressed gene is set to be TPM/FPKM>1

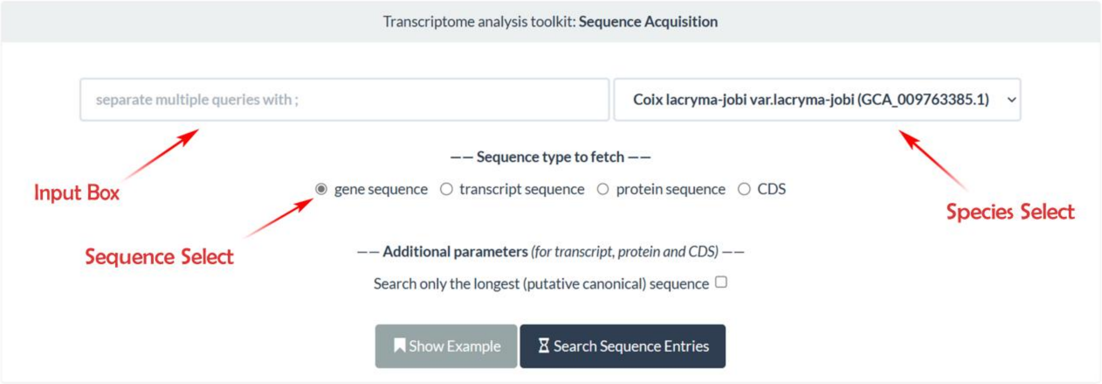
Obtain sequences of the target gene based on genome annotation profiles.

Preview or download raw gene expression data table used to generate the plots.

Figure 3.1.2 Expression Plots and Fundamental Info

3.2 Raw Sequence Acquisition

On this page, you have the option to swiftly retrieve gene, transcript, or protein sequences associated with your chosen Gene ID. Simply enter the Gene ID of interest into the query box. If you wish to query multiple Gene IDs simultaneously, please separate them using a semicolon ";".


Transcriptome analysis toolkit: Sequence Acquisition
separate multiple queries with ;
Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)
Input Box
Sequence Select
Species Select
Show Example
Search Sequence Entries

—— Sequence type to fetch ——

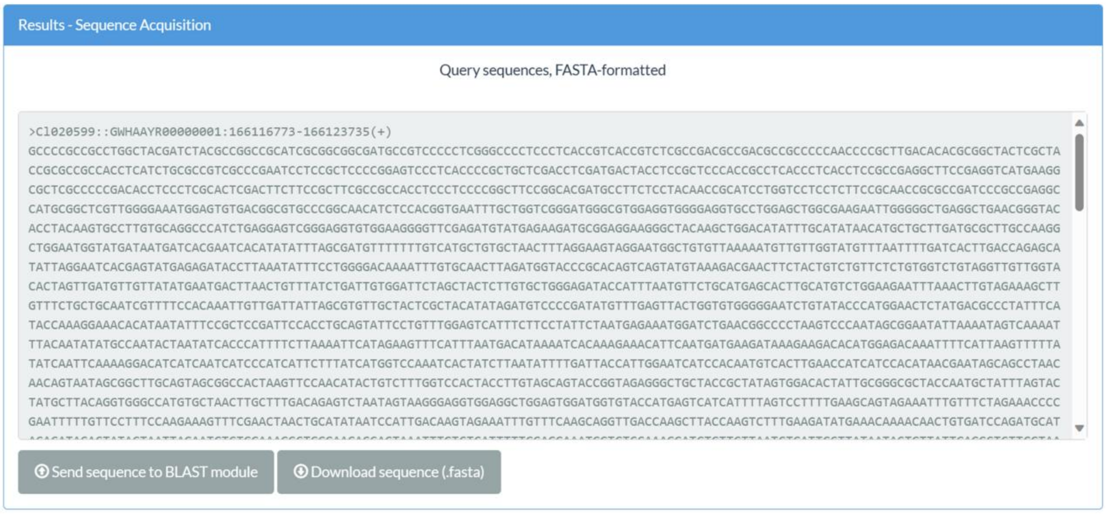
● gene sequence ○ transcript sequence ○ protein sequence ○ CDS

—— Additional parameters *(for transcript, protein and CDS)* ——

Search only the longest (putative canonical) sequence □

Figure 3.2.1 original interface of Raw Sequence Acquisition

After a short wait of a few seconds, you will find the FASTA-formatted sequences displayed below on the page. There's no need to copy the sequences manually. Simply click the "**Send sequence to BLAST module**" button, and these sequences will automatically populate the BLAST window for you.



Results - Sequence Acquisition

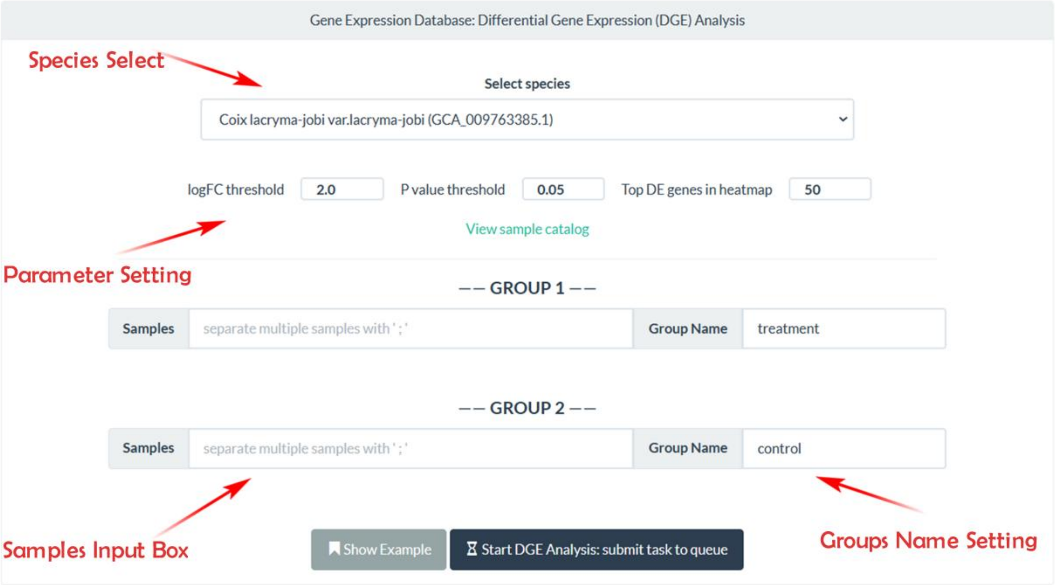
Query sequences, FASTA-formatted

```
>Cl020599::GWHAAYR00000001:166116773-166123735(+)
GCCCCGCCGCCTGGCTACGATCTACGCCGGCCGCATCGCGGCGGCGATGCCGTCCCCCTCGGGCCCCTCCCTCACCGTCACCGTCTCGCCGACGCCGACGCCGCCCCCAACCCCGCTTGACACACGCGGCTACTCGCTA
CCGCGCCGCCACCTCATCTGCGCCGTCGCCCGAATCCTCCGCTCCCCGGAGTCCCTCACCCCGCTGCTCGACCTCGATGACTACCTCCGCTCCCACCGCCTCACCCTCACCTCCGCCGAGGCTTCCGAGGTCATGAAGG
CGCTCGCCCCCGACACCTCCCTCGCACTCGACTTCTTCCGCTTCGCCGCCACCTCCCTCCCCGGCTTCCGGCACGATGCCTTCTCCTACAACCGCATCCTGGTCCTCCTCTTCCGCAACCGCGCCGATCCCGCCGAGGC
CATGCGGCTCGTTGGGGAAATGGAGTGTGACGGCGTGCCCGGCAACATCTCCACGGTGAATTTGCTGGTCGGGATGGGCGTGGAGGTGGGGAGGTGCCTGGAGCTGGCGAAGAATTGGGGGCTGAGGCTGAACGGGTAC
ACCTACAAGTGCCTTGTGCAGGCCCATCTGAGGAGTCGGGAGGTGTGGAAGGGGTTCGAGATGTATGAGAAGATGCGGAGGAAGGGCTACAAGCTGGACATATTTGCATATAACATGCTGCTTGATGCGCTTGCCAAGG
CTGGAATGGTATGATAATGATCACGAATCACATATATTTAGCGATGTTTTTTTGTCATGCTGTGCTAACTTTAGGAAGTAGGAATGGCTGTGTTAAAAATGTTGTTGGTATGTTTAATTTTGATCACTTGACCAGAGCA
TATTAGGAATCACGAGTATGAGAGATACCTTAAATATTTCCTGGGGACAAAATTTGTGCAACTTAGATGGTACCCGCACAGTCAGTATGTAAAGACGAACTTCTACTGTCTGTTCTCTGTGGTCTGTAGGTTGTTGGTA
CACTAGTTGATGTTGTTATATGAATGACTTAACTGTTTATCTGATTGTGGATTCTAGCTACTCTTGTGCTGGGAGATACCATTTAATGTTCTGCATGAGCACTTGCATGTCTGGAAGAATTTAAACTTGTAGAAAGCTT
GTTTCTGCTGCAATCGTTTTCCACAAATTGTTGATTATTAGCGTGTTGCTACTCGCTACATATAGATGTCCCCGATATGTTTGAGTTACTGGTGTGGGGGAATCTGTATACCCATGGAACTCTATGACGCCCTATTTCA
TACCAAAGGAAACACATAATATTTCCGCTCCGATTCCACCTGCAGTATTCCTGTTTGGAGTCATTTCTTCCTATTCTAATGAGAAATGGATCTGAACGGCCCCTAAGTCCCAATAGCGGAATATTAAAATAGTCAAAAT
TTACAATATATGCCAATACTAATATCACCCATTTTCTTAAAATTCATAGAAGTTTCATTTAATGACATAAAATCACAAAGAAACATTCAATGATGAAGATAAAGAAGACACATGGAGACAAATTTTCATTAAGTTTTTA
TATCAATTCAAAAGGACATCATCAATCATCCCATCATTCTTTATCATGGTCCAAATCACTATCTTAATATTTTGATTACCATTGGAATCATCCACAATGTCACTTGAACCATCATCCACATAACGAATAGCAGCCTAAC
AACAGTAATAGCGGCTTGCAGTAGCGGCCACTAAGTTCCAACATACTGTCTTTGGTCCACTACCTTGTAGCAGTACCGGTAGAGGGCTGCTACCGCTATAGTGGACACTATTGCGGGCGCTACCAATGCTATTTAGTAC
TATGCTTACAGGTGGGCCATGTGCTAACTTGCTTTGACAGAGTCTAATAGTAAGGGAGGTGGAGGCTGGAGTGGATGGTGTACCATGAGTCATCATTTTAGTCCTTTTGAAGCAGTAGAAATTTGTTTCTAGAAACCCC
GAATTTTTGTTCCTTTCCAAGAAAGTTTCGAACTAACTGCATATAATCCATTGACAAGTAGAAATTTGTTTCAAGCAGGTTGACCAAGCTTACCAAGTCTTTGAAGATATGAAACAAAACAACTGTGATCCAGATGCAT
```

Figure 3.2.2 results interface of Raw Sequence Acquisition

3.3 Differential Gene Expression(DGE) Analysis

Once you have obtained gene expression data of interest, your next step might involve investigating the variations in expression across different tissues or organs. To assist with this, we offer a comprehensive platform for Differential Gene Expression (DGE) Analysis.



Gene Expression Database: Differential Gene Expression (DGE) Analysis

Species Select

Select species: Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)

logFC threshold: 2.0

P value threshold: 0.05

Top DE genes in heatmap: 50

Parameter Setting

—— GROUP 1 ——

Samples: separate multiple samples with ' ; '

Group Name: treatment

—— GROUP 2 ——

Samples: separate multiple samples with ' ; '

Group Name: control

Samples Input Box

Groups Name Setting

Figure 3.3.1 the interface of Differential Gene Expression Analysis

Here's a step-by-step guide to help you navigate the process:

- **Select the species to which your genes belong.**
- **Adjust the relevant parameters until they are appropriate (default settings are also available).**
- **Enter the series of samples into the provided input box, separating them with**

semicolons.

- **Assign brief names to the groups for easier identification in the subsequent results.**
- **Initiate the analysis by clicking the "Start" button.**

If you are using this function for the first time, you can click on the "Show Example" button. This will automatically populate the input box with a set of prepared sample sequences.

After waiting for approximately 2 minutes, a comprehensive set of analysis results will be presented. These results have been categorized into four distinct sections for ease of comprehension.

Project Summary n this section, we provide an overview of the species, groups, and sample accessions you have selected for analysis. Additionally, we have compiled a table in CSV format that highlights the differentially expressed genes between groups, accompanied by detailed parameters.

Overview & Normalization of Expression Data This section includes three figures that elucidate the expression data provided. The first figure features raw and normalized sample TPM (Transcripts Per Million) boxplots, showcasing the distribution before and after normalization for each sample. The second figure depicts density curves of gene expression across samples, offering insight into the overall expression trends of genes. The third figure presents a two-dimensional correlation heatmap of expression among samples, using varying shades of color to visually represent the correlation coefficients between samples.

Principal Component Analysis In this section, we offer a Principal Component Analysis (PCA) plot. This plot aids in intuitively grasping the similarity between samples. For result accuracy, ellipses are excluded when the number of samples is insufficient.

Differential Analysis his part features a classical volcano plot, which effectively displays the differential expression status between groups. By referring to the provided key, you can readily understand the distribution of genes categorized as up-regulated, down-regulated, or not significantly altered. Subsequently, a Heatmap showcasing the most differentially expressed genes (based on TPM) illustrates the overall expression changes across multiple samples.

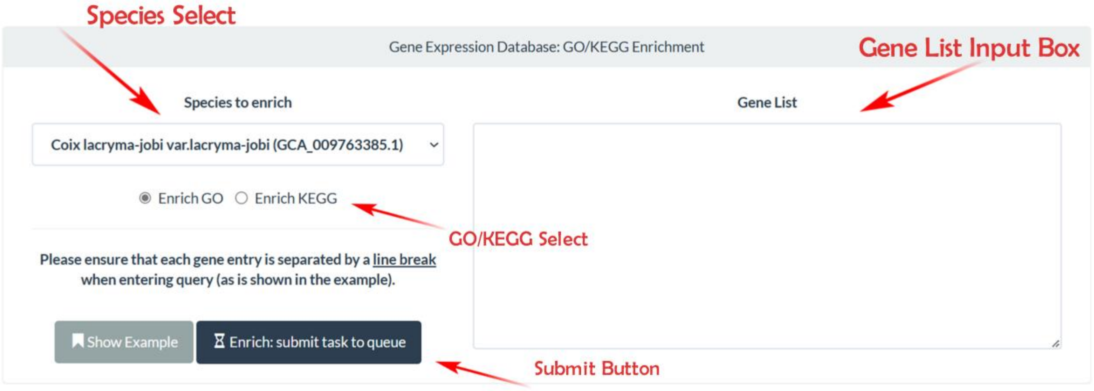
Preview and download options are made available. You can simply click the corresponding button beneath each project to preview or download the results as needed.

3.4 GO/KEGG Enrichment

We offer a user-friendly tool for conducting GO/KEGG enrichment analysis. This tool enables you to explore functional annotation details of gene transcripts in representative Andropogoneae species. To proceed, simply follow these steps:

- **Gather the gene IDs you're interested in.**
- **Enter the gene IDs in the designated Gene List box (each ID separated by a line break).**

- **Select the enrichment method you need (GO/KEGG).**
- **Click the Enrich button to submit your request.**


Species Select
Gene Expression Database: GO/KEGG Enrichment
Gene List Input Box
Species to enrich
Gene List
Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)
● Enrich GO ○ Enrich KEGG
GO/KEGG Select
Show Example
Enrich: submit task to queue
Submit Button

Please ensure that each gene entry is separated by a line break when entering query (as is shown in the example).

Figure 3.4.1 interface of GO/KEGG Enrichment

If you are using this function for the first time, you can click the "**Show Example**" button. A selection of gene lists that we have prepared will be automatically displayed in the left box. (Currently, we have included effective data and an example from *Zea mays.*)

Once the analysis is complete, a figure and a table will be presented. The figure is an interactive bubble plot where the color of each bubble corresponds to its p-value, and the size of the bubble represents the number of genes it contains. The ontology associated with each bubble is displayed on the right side of the axis. Hover your cursor over any bubble to view detailed information about it. Additionally, you can click the button below to download a static image of the plot.

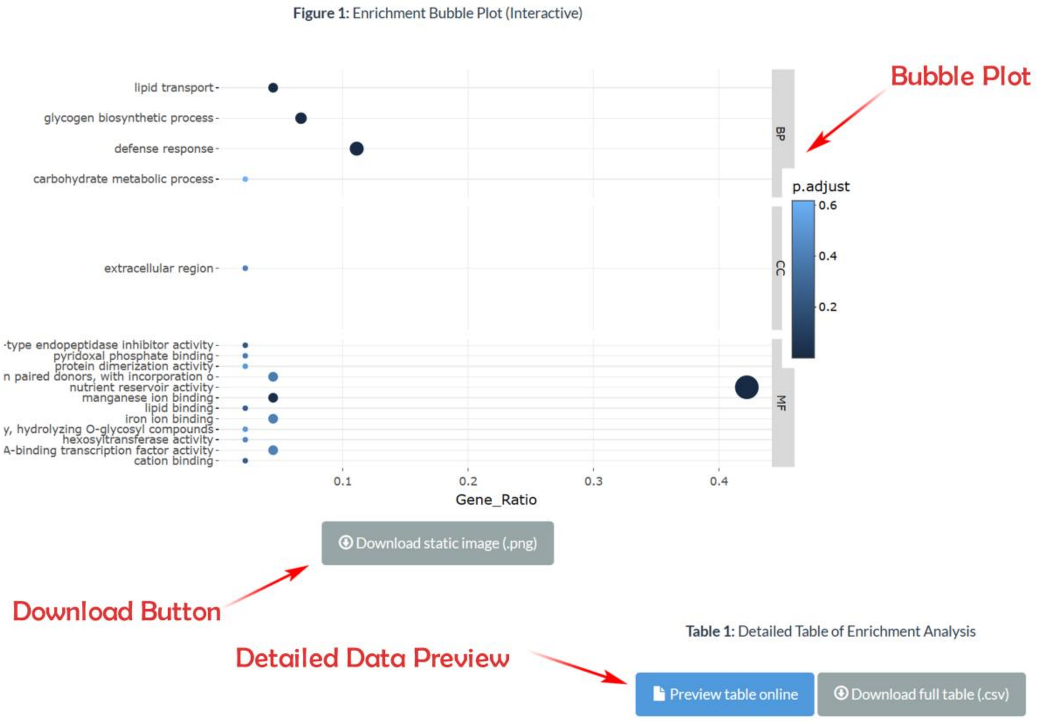


Figure 1: Enrichment Bubble Plot (Interactive)

Table 1: Detailed Table of Enrichment Analysis

Bubble Plot
lipid transport
glycogen biosynthetic process
defense response
carbohydrate metabolic process
extracellular region
-type endopeptidase inhibitor activity
pyridoxal phosphate binding
protein dimerization activity
n paired donors, with incorporation o
nutrient reservoir activity
manganese ion binding
lipid binding
iron ion binding
y, hydrolyzing O-glycosyl compounds
hexosyltransferase activity
A-binding transcription factor activity
cation binding
BP
CC
MF
p.adjust
0.6
0.4
0.2
0.1
0.2
0.3
0.4
Gene_Ratio
Download static image (.png)
Download Button
Detailed Data Preview
Preview table online
Download full table (.csv)

Figure 3.4.2 bubble plot of GO/KEGG Enrichment result

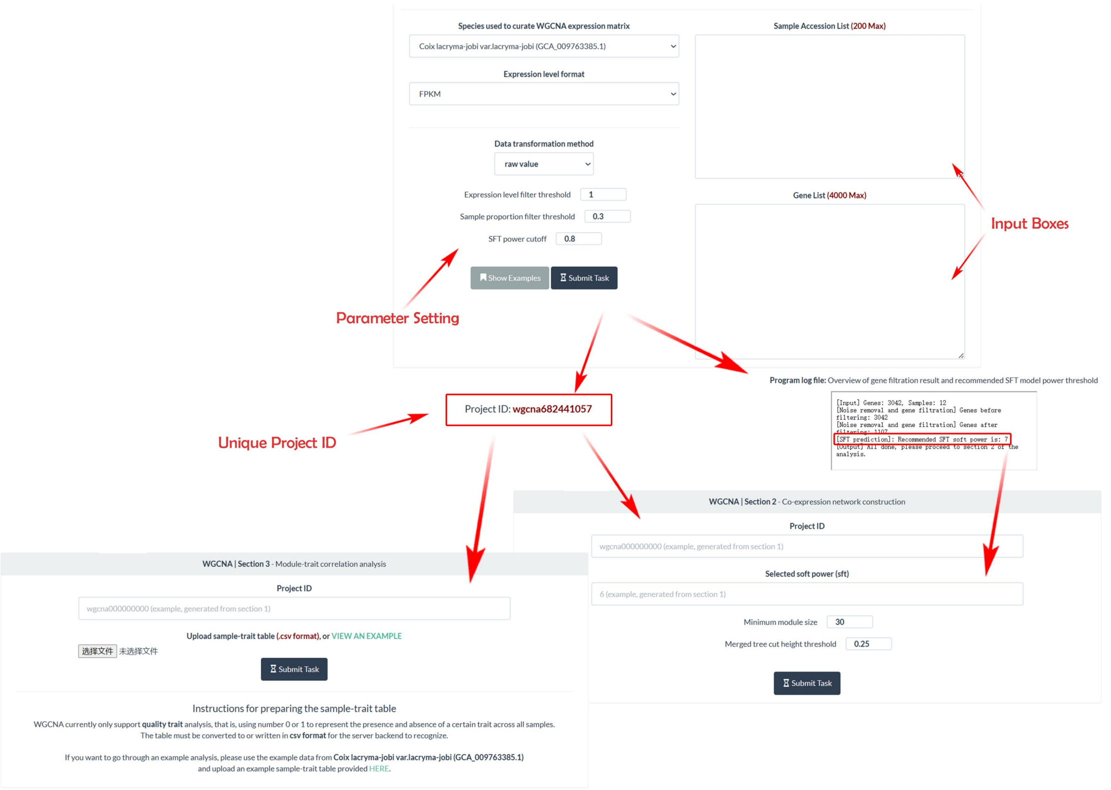
The table contains comprehensive data from the Enrichment Analysis, and it includes options for previewing and downloading the data.

3.5 Weighted Gene Co-expression Network Analysis(WGCNA)

Weighted Gene Co-expression Network Analysis (WGCNA) is a commonly used analytical method in systems biology and bioinformatics. This method is based on the expression patterns of genes across different samples, constructing a co-expression network among genes. By calculating the correlation between pairs of genes, highly correlated genes are grouped into modules. Each module represents a set of genes with similar expression patterns across samples. This approach not only focuses on the expression changes of individual genes but also emphasizes the coordinated regulatory relationships among sets of genes. It helps researchers gain a better understanding of the relationship between genes and phenotypes.

We have divided the overall WGCNA process into three sections. Section 1 covers Data curation, filtering, and sft selection. Section 2 involves Co-expression network construction, and Section 3 focuses on Module-trait correlation analysis. It's noteworthy that the analyses in Section 2 and Section 3 are relatively independent, as you will observe from the subsequent workflow. The detailed operational steps are as follows:

- **Navigate to the Section 1 interface, select the species you wish to analyze, and input the Sample Accession List and Gene List into the corresponding fields after adjusting relevant parameters**
- **Click the "Submit Task" button to initiate the analysis**
- **After the initial filtering, you will receive a unique Project ID. This ID will allow you to view completed analysis results at any time from the Result Viewer. The Section 1 results page includes a Program log file, where the "SFT prediction" will assist you in Section 2 analysis.**
- **Switch to the Section 2 page, input the Project ID and Recommended SFT soft power into the corresponding fields, and click the button to perform the analysis**
- **Switch to the Section 3 page, input the Project ID into the corresponding field, and click the button to initiate the analysis**



Species used to curate WGCNA expression matrix: Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)

Expression level format: FPKM

Data transformation method: raw value

Expression level filter threshold: 1

Sample proportion filter threshold: 0.3

SFT power cutoff: 0.8

Sample Accession List (200 Max)

Gene List (4000 Max)

Project ID: wgcna682441057

Program log file: Overview of gene filtration result and recommended SFT model power threshold

```
[Input] Genes: 3042, Samples: 12
[Noise removal and gene filtration] Genes before
filtering: 3042
[Noise removal and gene filtration] Genes after
filtering: 1107
[SFT prediction]: Recommended SFT soft power is: 7
[Output] All done, please proceed to section 2 of the
analysis.
```

WGCNA | Section 2 - Co-expression network construction

Project ID: wgcna000000000 (example, generated from section 1)

Selected soft power (sft): 6 (example, generated from section 1)

Minimum module size: 30

Merged tree cut height threshold: 0.25

WGCNA | Section 3 - Module-trait correlation analysis

Project ID: wgcna000000000 (example, generated from section 1)

Upload sample-trait table (.csv format), or VIEW AN EXAMPLE

未选择文件

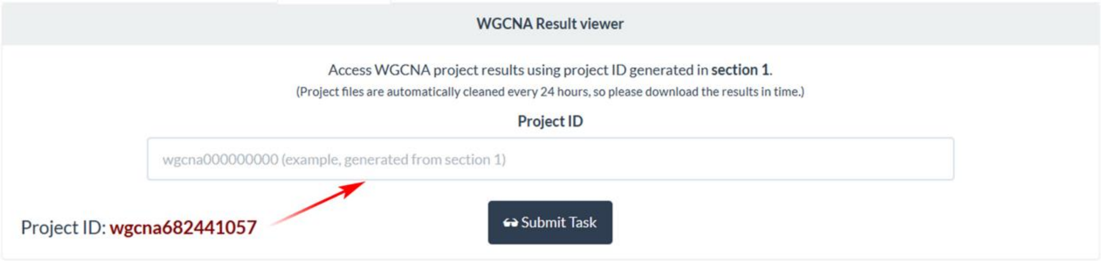
Instructions for preparing the sample-trait table

WGCNA currently only support **quality trait** analysis, that is, using number 0 or 1 to represent the presence and absence of a certain trait across all samples. The table must be converted to or written in **csv format** for the server backend to recognize.

If you want to go through an example analysis, please use the example data from **Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)** and upload an example sample-trait table provided HERE.

Figure 3.5.1 interface and flow path of WGCNA

After all analyses are completed, you can open the Result Viewer interface at any time. Enter the Project ID obtained during the analysis to quickly view or download the analysis results. The entire process will generate four figures: Sample cluster dendrogram, Scale independence and mean connectivity plot (from section 1), Network and cluster dendrogram (from section 2), and Module-trait correlation heatmap (from section 3), along with the associated tables.



WGCNA Result viewer

Access WGCNA project results using project ID generated in **section 1**.

(Project files are automatically cleaned every 24 hours, so please download the results in time.)

Project ID: wgcna000000000 (example, generated from section 1)

Project ID: wgcna682441057

Figure 3.5.2 interface of Result Viewer

Chapter 4

4 Toolkit III: Gene Family

4.1 Family Identification(HMM & BLASTP)

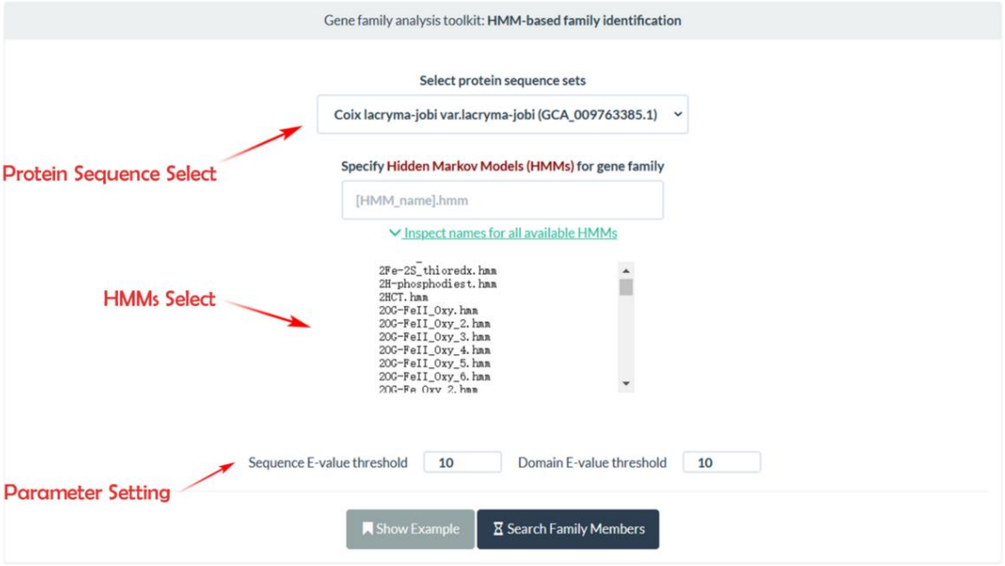
In most cases, genes belonging to the same gene family exhibit noticeable similarities in both structure and function. To cater to your requirements for identifying shared domains and distinctions within the gene sequences of your interest, we offer two distinct gene sequence alignment methods. Below, you will find a concise overview of these methods.

Hidden Markov Model(HMM) A statistical modeling method used for predicting gene sequences based on known ones. Utilizing training data from established gene sequences, it identifies common structural and feature traits within gene families. This allows the creation of gene models to predict unknown gene sequences.

BLASTP A protein sequence alignment method that identifies similarities between unknown protein sequences and a pre-prepared protein database. It assigns scores to assess the level of similarity, indicating the likelihood of a new gene when a higher score is achieved.

Identifying gene families requires careful consideration of multiple factors, as relying on single methods can lead to errors. To enhance the accuracy and reliability of gene identification, cross-validation is recommended. Follow these steps for improved results:

- **Select the desired protein sequences (separated by species).**
- **From the HMMs menu, select the relevant gene model.**
- **Adjust parameters as needed (defaults are provided).**
- **Click the "Search Family Members" button.**



Gene family analysis toolkit: **HMM-based family identification**

Select protein sequence sets

Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)

Specify Hidden Markov Models (HMMs) for gene family

[HMM_name].hmm

Inspect names for all available HMMs

```
2Fe-2S_thioredx.hmm
2H-phosphodiest.hmm
2HCT.hmm
2OG-FeII_Oxy.hmm
2OG-FeII_Oxy_2.hmm
2OG-FeII_Oxy_3.hmm
2OG-FeII_Oxy_4.hmm
2OG-FeII_Oxy_5.hmm
2OG-FeII_Oxy_6.hmm
2OG-Fe_Oxy_2.hmm
```

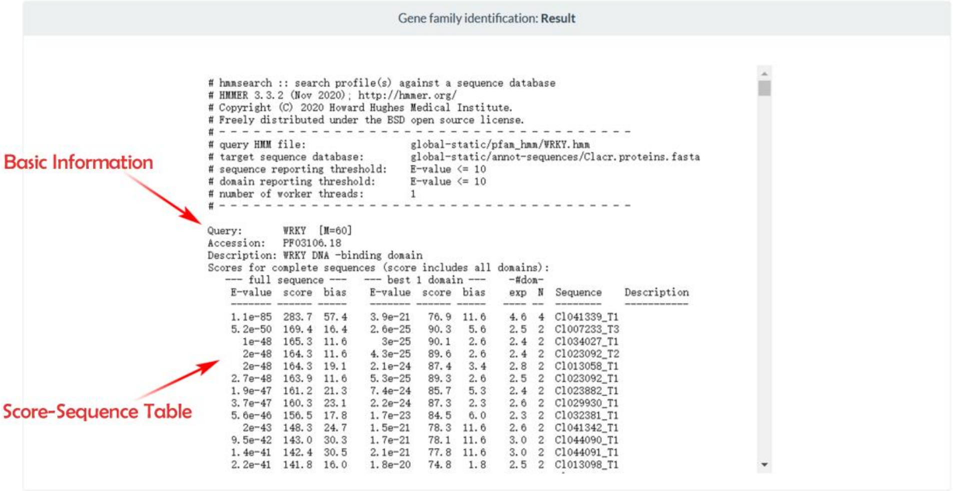
Sequence E-value threshold: 10

Domain E-value threshold: 10

Figure 4.1.1 interface of HMM-based Family Identification

If you are using this function for the first time, you can click on the "**Show Examples**" button. This will automatically populate the input box with a preset model that we have provided.

The HMM-based results include a score-sequence table, which is organized from highest to lowest score, indicating the similarity of each potential gene with the specific model. Additionally, comprehensive domain annotations for each sequence and alignments for individual domains are presented. The identification process concludes with a statistical summary.

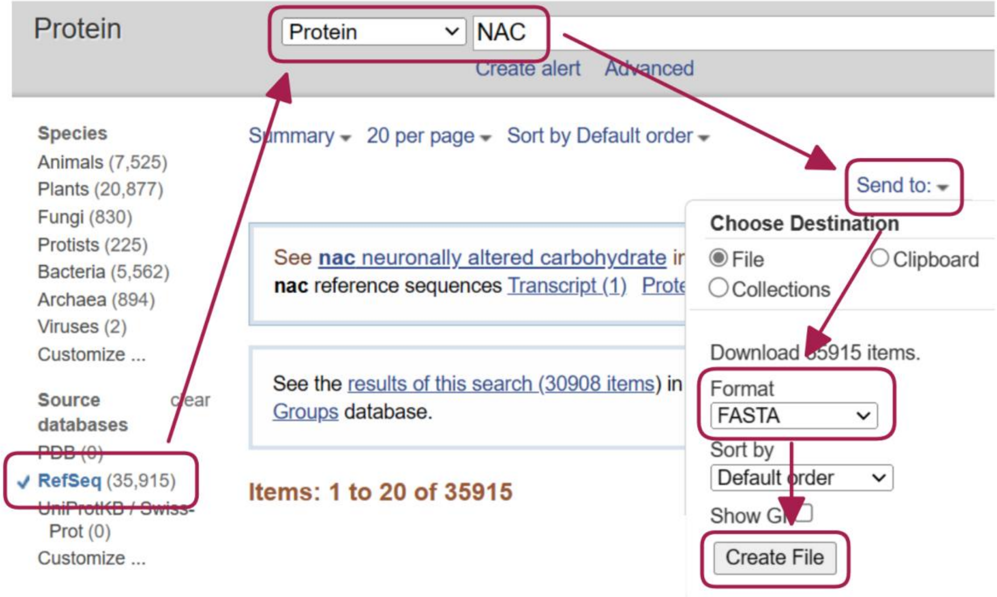


```
# hmmsearch :: search profile(s) against a sequence database
# HMMER 3.3.2 (Nov 2020); http://hmmer.org/
# Copyright (C) 2020 Howard Hughes Medical Institute.
# Freely distributed under the BSD open source license.
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# query HMM file:                  global-static/pfam_hmm/WRKY.hmm
# target sequence database:        global-static/annot-sequences/Clacr.proteins.fasta
# sequence reporting threshold:    E-value <= 10
# domain reporting threshold:      E-value <= 10
# number of worker threads:        1
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Query:       WRKY  [M=60]
Accession:   PF03106.18
Description: WRKY DNA -binding domain
Scores for complete sequences (score includes all domains):
   --- full sequence ---   --- best 1 domain ---    -#dom-
    E-value  score  bias    E-value  score  bias    exp  N  Sequence     Description
    ------- ------ -----    ------- ------ -----   ---- --  --------     -----------
    1.1e-85  283.7  57.4    3.9e-21   76.9  11.6    4.6  4  Cl041339_T1
    5.2e-50  169.4  16.4    2.6e-25   90.3   5.6    2.5  2  Cl007233_T3
      1e-48  165.3  11.6      3e-25   90.1   2.6    2.4  2  Cl034027_T1
      2e-48  164.3  11.6    4.3e-25   89.6   2.6    2.4  2  Cl023092_T2
      2e-48  164.3  19.1    2.1e-24   87.4   3.4    2.8  2  Cl013058_T1
    2.7e-48  163.9  11.6    5.3e-25   89.3   2.6    2.5  2  Cl023092_T1
    1.9e-47  161.2  21.3    7.4e-24   85.7   5.3    2.4  2  Cl023882_T1
    3.7e-47  160.3  23.1    2.2e-24   87.3   2.3    2.6  2  Cl029930_T1
    5.6e-46  156.5  17.8    1.7e-23   84.5   6.0    2.3  2  Cl032381_T1
      2e-43  148.3  24.7    1.5e-21   78.3  11.6    2.6  2  Cl041342_T1
    9.5e-42  143.0  30.3    1.7e-21   78.1  11.6    3.0  2  Cl044090_T1
    1.4e-41  142.4  30.5    2.1e-21   77.8  11.6    3.0  2  Cl044091_T1
    2.2e-41  141.8  16.0    1.8e-20   74.8   1.8    2.5  2  Cl013098_T1
```

Figure 4.1.2 A example of result based on HMM

Regarding the outcome derived from BLASTP analysis, it is recommended to acquire a representative sequence (RefSeq) from a protein family to identify members in different species. The RefSeq can be obtained through manual curation or downloaded from a publicly accessible RefSeq database.



See **nac** neuronally altered carbohydrate ir

nac reference sequences Transcript (1) Prote

See the results of this search (30908 items) in Groups database.

Items: 1 to 20 of 35915

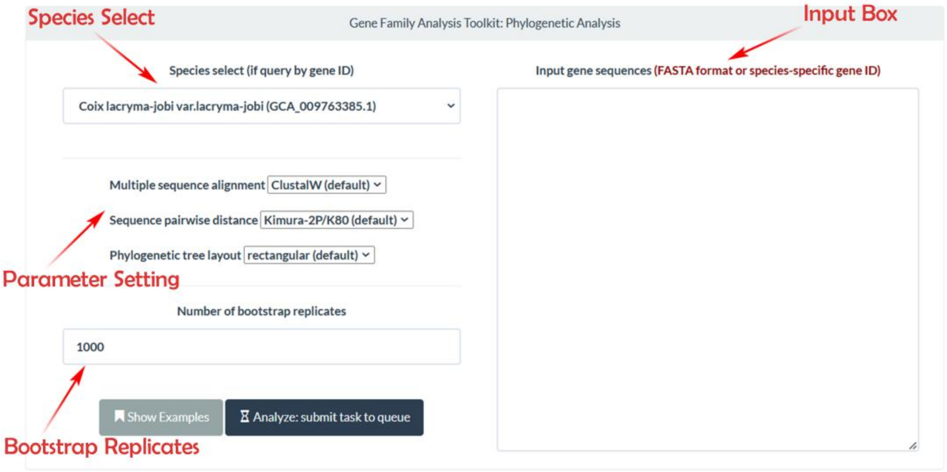
Figure 4.1.3 guide for downloading RefSeq from the NCBI database

To retrieve the relevant outcomes, kindly insert the preferred protein sequences within the designated input field. Following this, select the appropriate RefSeq, make adjustments to the advanced parameters (set to default if preferred), and proceed by clicking on the **BLAST** button. Subsequently, the comprehensive alignment details for the sequences will be presented in a systematic manner.

4.2 Phylogenetic Analysis

Once you have identified a set of potential genes through cross-validation, creating a gene family phylogenetic tree can help elucidate the evolutionary and genetic connections among these genes. To facilitate this, we offer a phylogenetic analysis toolkit for the automatic construction of a foundational tree. The recommended procedure involves the following steps:

- Select the desired species.
- Configure appropriate parameters and bootstrap replicates (default values are provided).
- Input the genes using either FASTA format or Gene IDs, obtained from your prior analysis.
- Click the "**Analyze**" button to submit the task to the queue.


Gene Family Analysis Toolkit: Phylogenetic Analysis

Species select (if query by gene ID)

Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)

Multiple sequence alignment: ClustalW (default)

Sequence pairwise distance: Kimura-2P/K80 (default)

Phylogenetic tree layout: rectangular (default)

Number of bootstrap replicates: 1000

Input gene sequences (FASTA format or species-specific gene ID)

Figure 4.2.1 the interface of Phylogenic Analysis

If you are using this function for the first time, you can click the "**Show Examples**" button. This will automatically populate the box with a set of preselected Gene IDs that we have prepared.

The outcome will present a preview of a phylogenetic tree, indicating the converted bootstrap value on each node. Generally, values above 50 indicate the reliability of this phylogenetic tree.

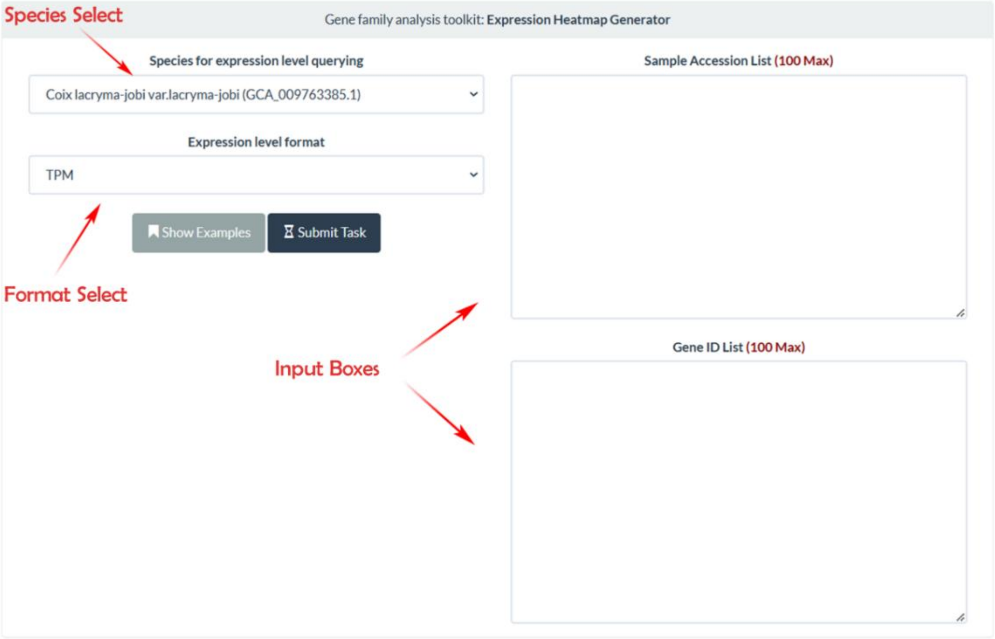
You have the option to download the static image in PNG format or obtain the editable tree in NWK format for importing into the MEGA (Molecular Evolutionary Genetics Analysis) software for further enhancement.

4.3 Family Expression Heatmap

For analyzing the expression levels of multiple gene families across various samples, we offer a heatmap generation tool designed to help you visualize your gene family data. Currently, this tool can handle a maximum of 100 sample accessions and 100 gene IDs for analysis. To utilize this tool, please follow these steps:

- **Select the species of your samples.**
- **Choose the preferred expression level format (defaulted to TPM).**
- **List your samples and gene IDs (up to 100) and input them in the provided boxes.**
- **Click the "Submit Task" button to initiate data submission.**

For first-time users, you can click the "Show Examples" button. This action will automatically populate the boxes with sample accessions and gene IDs from a set of prepared lists. (Currently, we have included effective data and an example from *Coix lacryma-jobi var. lacryma-jobi*.)



Gene family analysis toolkit: **Expression Heatmap Generator**

Species for expression level querying: Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)

Expression level format: TPM

Figure 4.3.1 interface of Gene Family Expression Heatmap Generator

Consequently, the system will display a visual representation comprising a heatmap depicting gene expression levels and a dendrogram illustrating sample clustering. This will be accompanied by a comprehensive table showcasing the gene expression data across various samples.

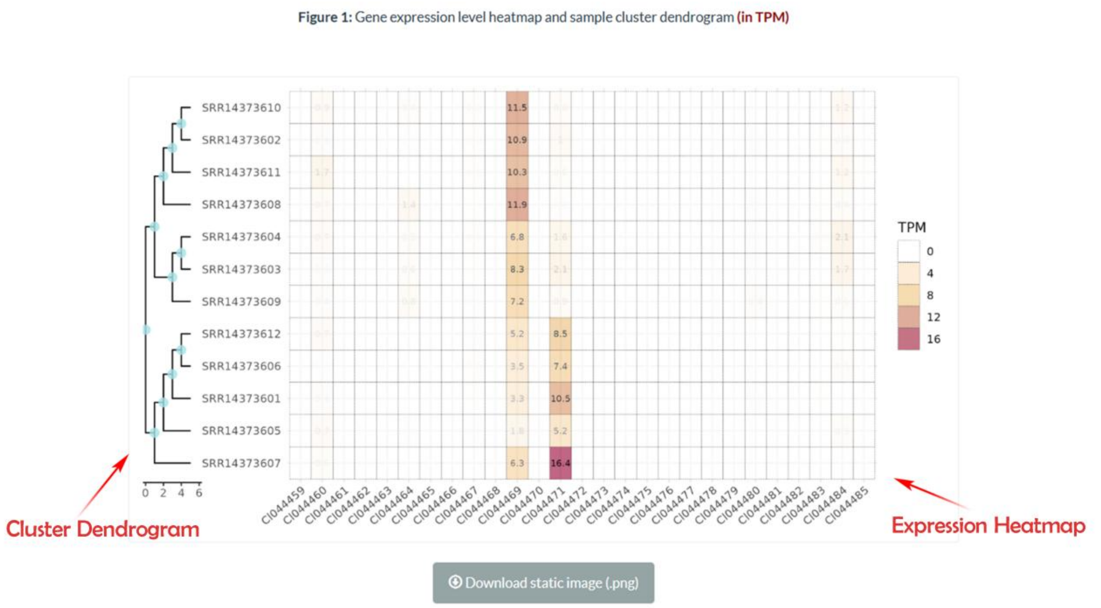


Figure 4.3.2 visualization of gene family expression

Within the interface, you'll notice that the intensity of colors on the right side signifies the relative expression level of each gene family. A darker shade indicates a higher expression for that gene family. The precise expression is quantified by the TPM value. On the left side, a cluster dendrogram has been constructed based on the gene family's expression pattern. This dendrogram visually represents the similarities and differences between gene families. For a more in-depth analysis, you can download the table we offer, which contains comprehensive data for your examination.

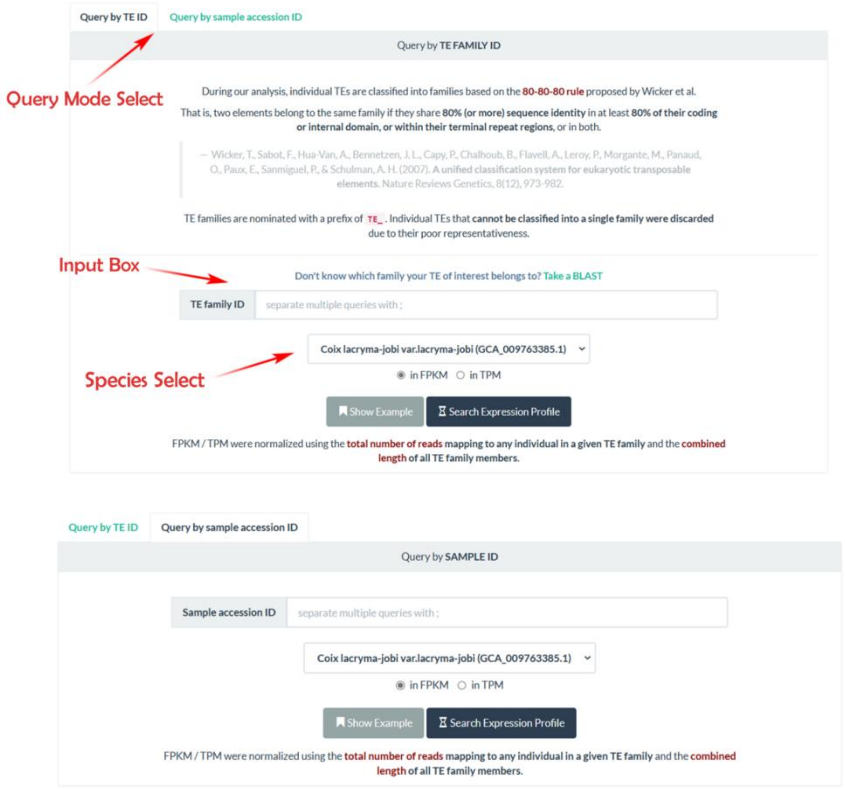
Chapter 5

5 Toolkit IV: TE Analysis

5.1 TE(Transposable Elements) Expression Query

TE expression data of representative Andropogoneae species has been analyzed and an interactive query interface has been provided. During our analysis, individual TEs are classified into families based on the 80-80-80 rule proposed by Wicker et al. That is, two elements belong to the same family if they share 80% (or more) sequence identity in at least 80% of their coding or internal domain, or within their terminal repeat regions, or in both. You can query any TEs you interested in this interface. Concrete steps we suggest are following:

- Select a suitable query mode
- Input TE ID/Sample Accession ID you interested in the box
- Choose corresponding species and the format you want
- Click **Search Expression Profile** button to query


Query by TE ID
Query by sample accession ID
Query Mode Select
Input Box
Species Select
TE family ID
separate multiple queries with ;
Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)
● in FPKM ○ in TPM
Show Example
Search Expression Profile
Query by TE ID
Query by sample accession ID
Sample accession ID
separate multiple queries with ;
Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)
● in FPKM ○ in TPM
Show Example
Search Expression Profile

Query by **TE FAMILY ID**

During our analysis, individual TEs are classified into families based on the **80-80-80 rule** proposed by Wicker et al.

That is, two elements belong to the same family if they share **80% (or more) sequence identity** in at least **80% of their coding or internal domain, or within their terminal repeat regions**, or in both.

> — Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., Sanmiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics, 8(12), 973-982.

TE families are nominated with a prefix of TE_. Individual TEs that **cannot be classified into a single family were discarded** due to their poor representativeness.

Don't know which family your TE of interest belongs to? Take a BLAST

FPKM / TPM were normalized using the **total number of reads mapping to any individual in a given TE family** and the **combined length of all TE family members.**

Query by **SAMPLE ID**

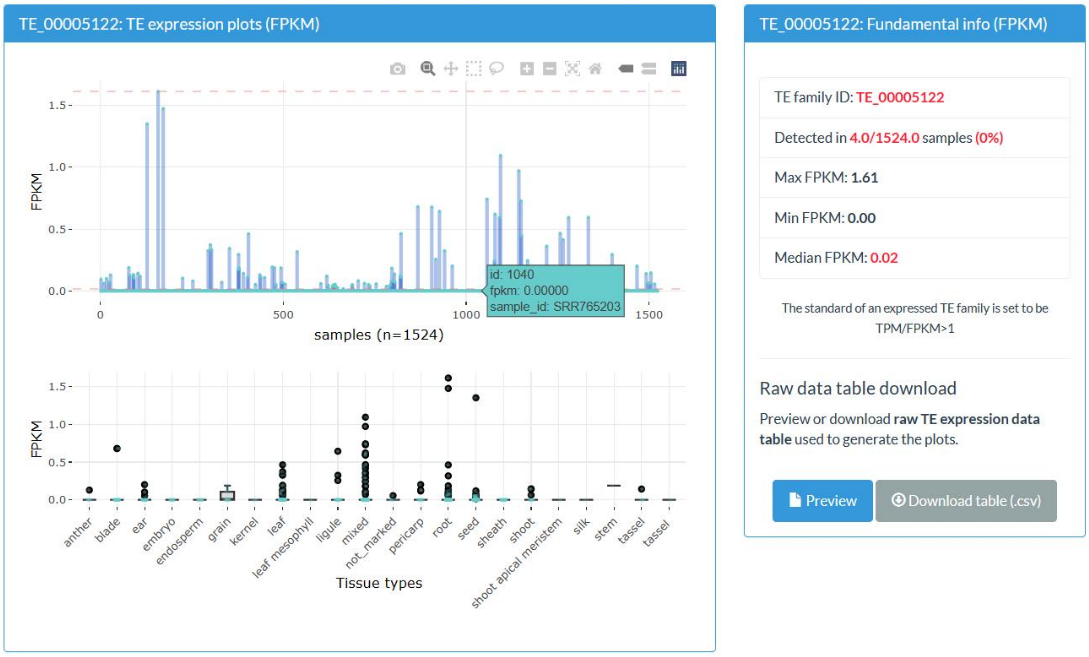
FPKM / TPM were normalized using the **total number of reads mapping to any individual in a given TE family** and the **combined length of all TE family members.**

Figure 5.1.1 interface of TE Expression Query

Waiting for a few seconds, you can see the results with two panels of the gene expression profiles for every TE ID.

Expression Plots We provide two interactive expression plots to display the profile results, both sample level and tissue level. In Sample level plots, stay your cursor on any sample data, then you can see its ID, value of TPM/FPKM and sample ID. Similarly, in tissue level, you can query the tissue type and its value of TPM/FPKM any sample data belongs to by the same way.

Fundamental Info This panel displays fundamental information regarding the query gene. Including frequency, max, min and median expression in samples, which give a basic recognition of your interested TE ID. The standard of an expressed TE family is set to be TPM/FPKM>1. Additional, buttons of previewing or downloading raw TE expression data table used to generate the plots are provided.



TE_00005122: Fundamental info (FPKM)

TE family ID: **TE_00005122**

Detected in **4.0/1524.0** samples **(0%)**

Max FPKM: **1.61**

Min FPKM: **0.00**

Median FPKM: **0.02**

The standard of an expressed TE family is set to be TPM/FPKM>1

Raw data table download

Preview or download **raw TE expression data table** used to generate the plots.

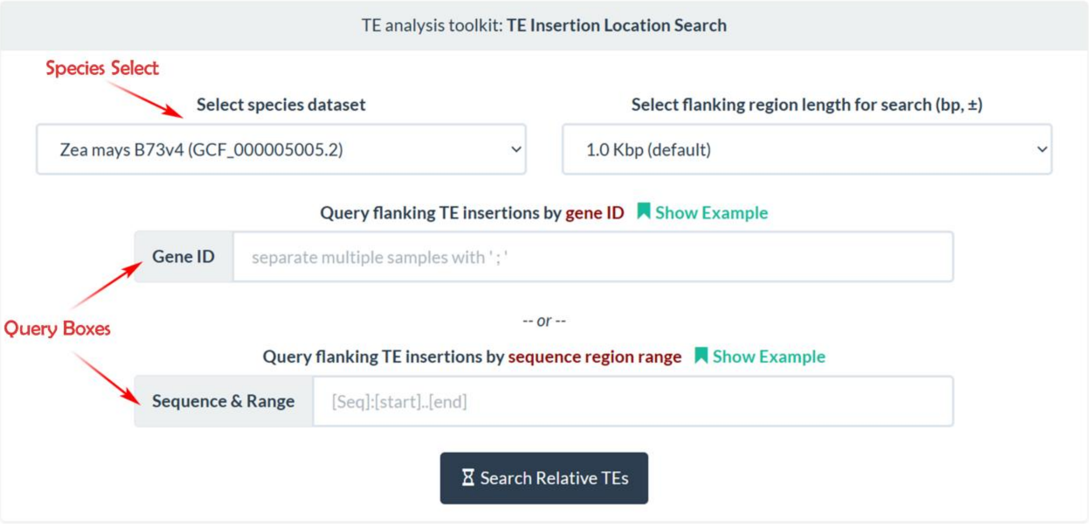
Figure 5.1.2 results of TE Expression Query

5.2 TE Insertion Location Search

TE(Transposable Element) is a type of DNA sequence that can move to different locations in the genome, and its insertion location in the genome is closely related to its biological function. To better understand the impact of TEs on genome structure and function, we have provided two methods for querying TE insertion locations: inputting Gene ID or sequence region range. These methods can help researchers identify where TEs are inserted, and further investigate their roles and biological functions in the genome. Concrete steps we suggest are following:

- Select the species you want to query
- Select suitable flanking region length(1.0 Kbp defaulted)
- Input the Gene ID or sequence region range you interested in the corresponding box
- Click **Search Relative TEs** button and wait for a minute

If you use this function first, you can click the **Show Examples** button. A preseted Gene ID/sequence region range will be filled in the corresponding box automatically.



TE analysis toolkit: **TE Insertion Location Search**

Select species dataset: Zea mays B73v4 (GCF_000005005.2)

Select flanking region length for search (bp, ±): 1.0 Kbp (default)

Query flanking TE insertions by gene ID

Gene ID: separate multiple samples with ' ; '

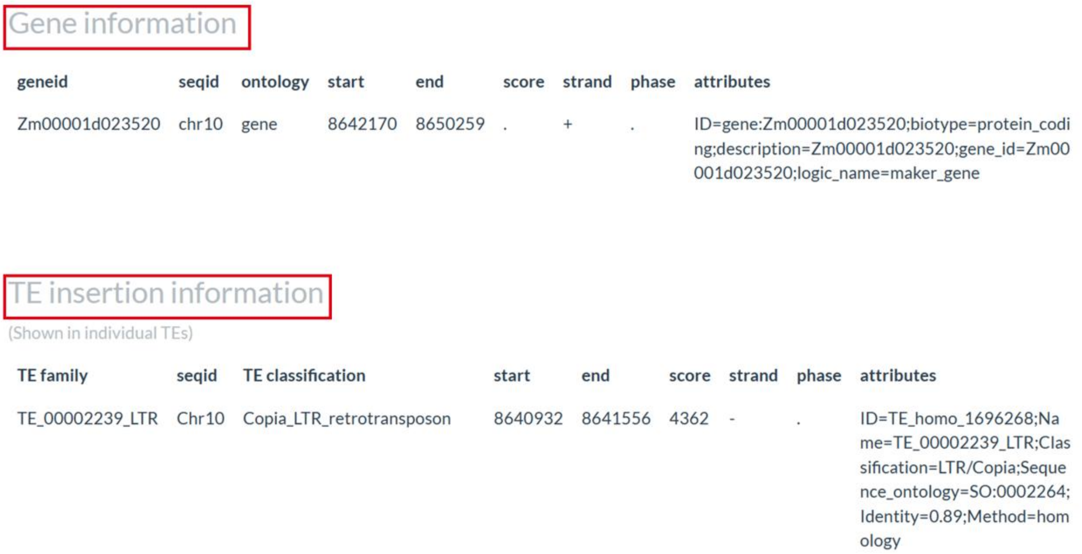
-- or --

Query flanking TE insertions by sequence region range

Sequence & Range: [Seq]:[start]..[end]

Figure 5.2.1 interface of TE Insertion Search

About the results, a table of the information of insertion location in what you query will be provided, including its classification, start and end, score, p-value and so on.



Gene information

geneid	seqid	ontology	start	end	score	strand	phase	attributes
Zm00001d023520	chr10	gene	8642170	8650259	.	+	.	ID=gene:Zm00001d023520;biotype=protein_coding;description=Zm00001d023520;gene_id=Zm00001d023520;logic_name=maker_gene

TE insertion information

(Shown in individual TEs)

TE family	seqid	TE classification	start	end	score	strand	phase	attributes
TE_00002239_LTR	Chr10	Copia_LTR_retrotransposon	8640932	8641556	4362	-	.	ID=TE_homo_1696268;Name=TE_00002239_LTR;Classification=LTR/Copia;Sequence_ontology=SO:0002264;Identity=0.89;Method=homology

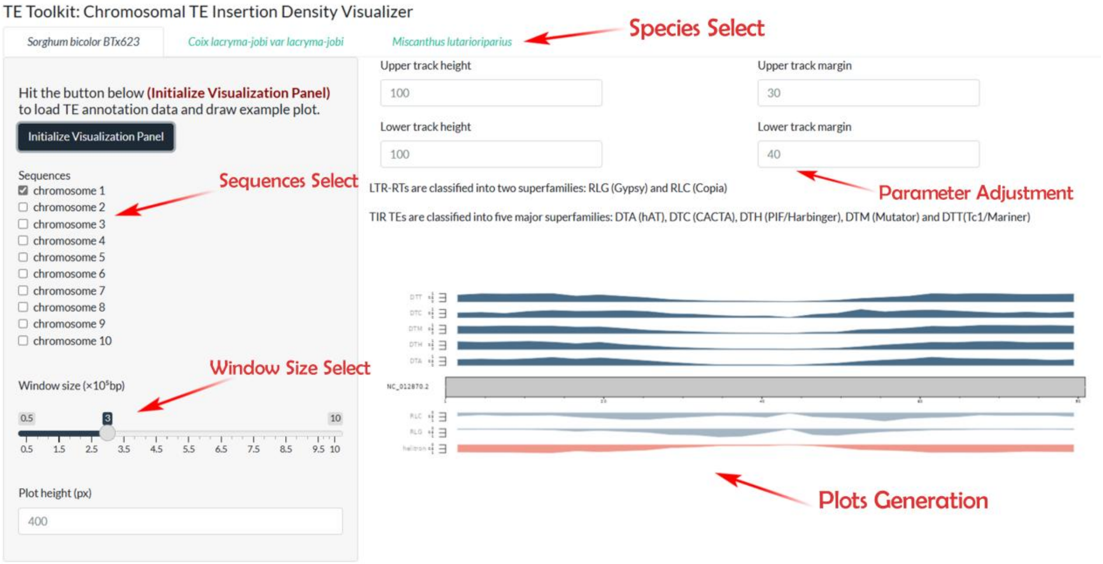
Figure 5.2.2 results of TE Insertion Search(query by Gene ID)

Additionally, if you query it by inputting Gene ID, its detailed information will be enclosed.

5.3 Chromosome-level TE Insertion Density Search

An interactive Chromosome level TE Insertion Density Search platform is provided, which calculates the number of TE on chromosomes to determine their distribution and study their role in genome structure and function. At the top, species can be selected. Select the desired chromosome on the left and adjust the relevant parameters on the right to obtain the corresponding TE annotation data and draw example plot. Sliding the left slider can adjust the visible window size. Concrete steps we suggest are following:

- Select the species you want to query
- Set suitable parameters and sequences you interested
- Click **Initialize Visualization Panel** button and wait for a minute



TE Toolkit: Chromosomal TE Insertion Density Visualizer

Sorghum bicolor BTx623 | *Coix lacryma-jobi var lacryma-jobi* | *Miscanthus lutarioriparius*

Species Select

Hit the button below **(Initialize Visualization Panel)** to load TE annotation data and draw example plot.

Initialize Visualization Panel

Sequences

☑ chromosome 1
☐ chromosome 2
☐ chromosome 3
☐ chromosome 4
☐ chromosome 5
☐ chromosome 6
☐ chromosome 7
☐ chromosome 8
☐ chromosome 9
☐ chromosome 10

Sequences Select

Window size (×10⁵bp): 3

Window Size Select

Plot height (px): 400

Upper track height: 100

Upper track margin: 30

Lower track height: 100

Lower track margin: 40

Parameter Adjustment

LTR-RTs are classified into two superfamilies: RLG (Gypsy) and RLC (Copia)

TIR TEs are classified into five major superfamilies: DTA (hAT), DTC (CACTA), DTH (PIF/Harbinger), DTM (Mutator) and DTT(Tc1/Mariner)

Plots Generation

Figure 5.3.1 interface of TE Insertion Density Visualizer(interactive)

About the results, a series of TE insertion density example plots will be generated. Defaulted window size is 3×10⁵ bp.

Chapter 6

6 Toolkit V: Epigenome Analysis

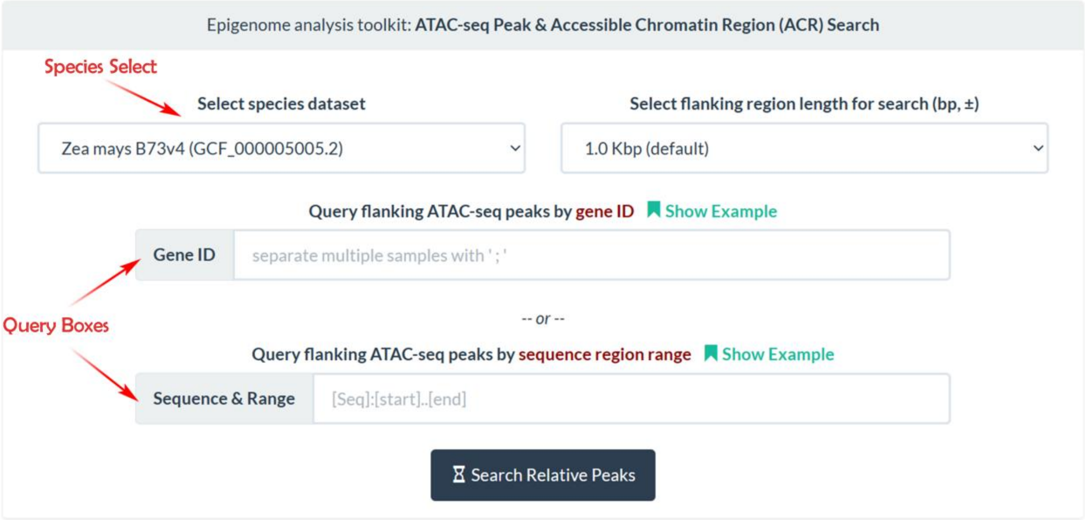
6.1 Chromatin Accessibility Search

Chromatin Accessibility Search is an analysis method that identifies open chromatin regions based on peak information from annotated ATAC-seq samples. By identifying open chromatin regions in the genome, this method helps to understand gene expression regulation mechanisms and gene regulatory events during cell differentiation and development. Two ways to search for open chromatin regions are provided: inputting a Gene ID from a certain species or a sequence range on a chromosome, both of which will return relevant information on all open chromatin regions at that location. Concrete steps we suggest are following:

- Select the species you want to query

- Select suitable flanking region length(1.0 Kbp defaulted)
- Input the Gene ID or sequence region range you interested in the corresponding box
- Click **Search Relative Peaks** button and wait for a minute

If you use this function first, you can click the **Show Examples** button. A preseted Gene ID/sequence region range will be filled in the corresponding box automatically.



Epigenome analysis toolkit: **ATAC-seq Peak & Accessible Chromatin Region (ACR) Search**

Select species dataset: Zea mays B73v4 (GCF_000005005.2)

Select flanking region length for search (bp, ±): 1.0 Kbp (default)

Query flanking ATAC-seq peaks by gene ID

Gene ID: separate multiple samples with ' ; '

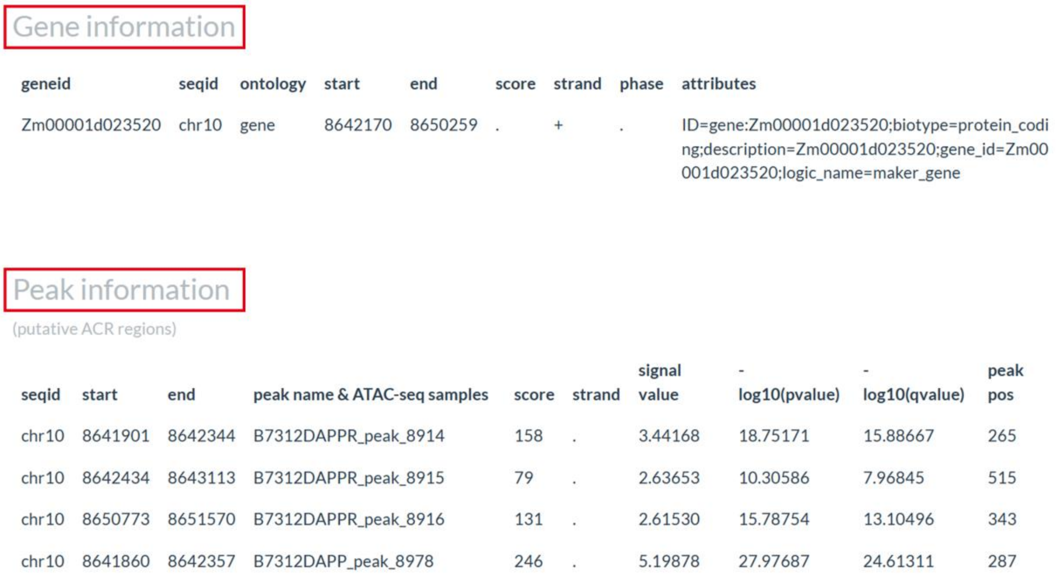
-- or --

Query flanking ATAC-seq peaks by sequence region range

Sequence & Range: [Seq]:[start]..[end]

Figure 6.1.1 interface of Chromatin Accessibility Search

About the results, a table of the information of all chromatin regions in what you query will be provided, including its start and end, score, p-value and so on.



Gene information

geneid	seqid	ontology	start	end	score	strand	phase	attributes
Zm00001d023520	chr10	gene	8642170	8650259	.	+	.	ID=gene:Zm00001d023520;biotype=protein_coding;description=Zm00001d023520;gene_id=Zm00001d023520;logic_name=maker_gene

Peak information

(putative ACR regions)

seqid	start	end	peak name & ATAC-seq samples	score	strand	signal value	-log10(pvalue)	-log10(qvalue)	peak pos
chr10	8641901	8642344	B7312DAPPR_peak_8914	158	.	3.44168	18.75171	15.88667	265
chr10	8642434	8643113	B7312DAPPR_peak_8915	79	.	2.63653	10.30586	7.96845	515
chr10	8650773	8651570	B7312DAPPR_peak_8916	131	.	2.61530	15.78754	13.10496	343
chr10	8641860	8642357	B7312DAPP_peak_8978	246	.	5.19878	27.97687	24.61311	287

Figure 6.1.2 results of Chromatin Accessibility Search(query by Gene ID)

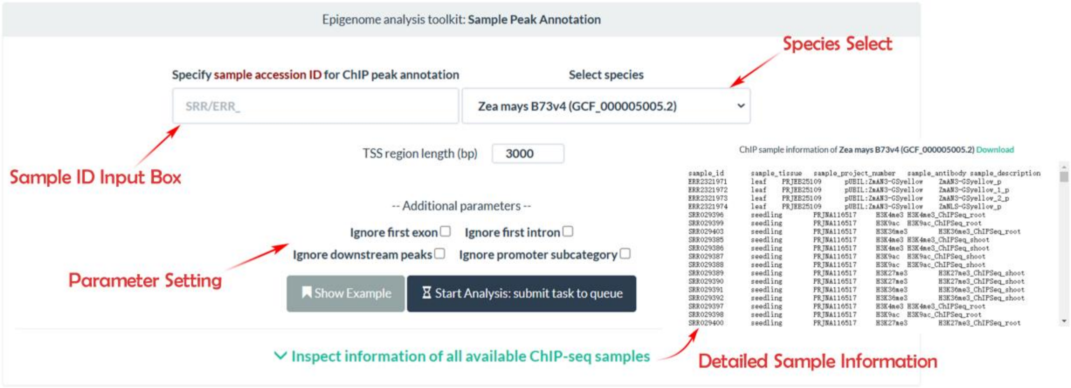
Additionally, if you query it by inputting Gene ID, its detailed information will be enclosed.

6.2 ChIP-seq Peak Annotation Analysis

We built a database for ChIP-Seq data of representative Andropogoneae species and streamlined the entire process of ChIP-Seq peak annotation analysis. Specify ChIP-Seq sample accession ID then you can get corresponding analysis results. Pull down the menu selection at the bottom of this interface. Then you can browse all available ChIP-Seq samples. It lists more detailed information of every sample so that can assist you in choosing the sample ID you interested. Concrete steps we suggest are following:

- Pull down the ChIP-Seq samples menu
- Choose a sample ID you interested then input it in the box
- Select the corresponding species
- Set TSS(Transcription Start Sequence) region length the results finally display(in unit of bp, default provided)
- Set additional parameters(not recommended)
- Click **Start Analysis** button and wait for a minute

If you use this function first, you can click the **Show Examples** button. A preseted sample accession ID will be filled in the boxes automatically.


Epigenome analysis toolkit: Sample Peak Annotation
Species Select
Sample ID Input Box
-- Additional parameters --
Parameter Setting
Show Example
Start Analysis: submit task to queue
Inspect information of all available ChIP-seq samples
Detailed Sample Information
Download

Specify sample accession ID for ChIP peak annotation: SRR/ERR_

Select species: Zea mays B73v4 (GCF_000005005.2)

TSS region length (bp): 3000

Ignore first exon □ Ignore first intron □
Ignore downstream peaks □ Ignore promoter subcategory □

ChIP sample information of Zea mays B73v4 (GCF_000005005.2)

```
sample_id	sample_tissue	sample_project_number	sample_antibody	sample_description
ERR2321971	leaf	PRJEB25109	pUBIL:ZmAN3-GSyellow	ZmAN3-GSyellow_p
ERR2321972	leaf	PRJEB25109	pUBIL:ZmAN3-GSyellow	ZmAN3-GSyellow_1_p
ERR2321973	leaf	PRJEB25109	pUBIL:ZmAN3-GSyellow	ZmAN3-GSyellow_2_p
ERR2321974	leaf	PRJEB25109	pUBIL:ZmAN3-GSyellow	ZmNLS-GSyellow_p
SRR029396	seedling	PRJNA116517	H3K4me3	H3K4me3_ChIPSeq_root
SRR029399	seedling	PRJNA116517	H3K9ac	H3K9ac_ChIPSeq_root
SRR029403	seedling	PRJNA116517	H3K36me3	H3K36me3_ChIPSeq_root
SRR029385	seedling	PRJNA116517	H3K4me3	H3K4me3_ChIPSeq_shoot
SRR029386	seedling	PRJNA116517	H3K4me3	H3K4me3_ChIPSeq_shoot
SRR029387	seedling	PRJNA116517	H3K9ac	H3K9ac_ChIPSeq_shoot
SRR029388	seedling	PRJNA116517	H3K9ac	H3K9ac_ChIPSeq_shoot
SRR029389	seedling	PRJNA116517	H3K27me3	H3K27me3_ChIPSeq_shoot
SRR029390	seedling	PRJNA116517	H3K27me3	H3K27me3_ChIPSeq_shoot
SRR029391	seedling	PRJNA116517	H3K36me3	H3K36me3_ChIPSeq_shoot
SRR029392	seedling	PRJNA116517	H3K36me3	H3K36me3_ChIPSeq_shoot
SRR029397	seedling	PRJNA116517	H3K4me3	H3K4me3_ChIPSeq_root
SRR029398	seedling	PRJNA116517	H3K9ac	H3K9ac_ChIPSeq_root
SRR029400	seedling	PRJNA116517	H3K27me3	H3K27me3_ChIPSeq_root
```

Figure 6.3.1 interface of ChIP-Seq Peak Annotation

The results of the analysis include five subgraphs:

Heatmap This subgraph displays the distance of every ChIP-Seq near the TSS region, every colored short line corresponds to a sequence. Typically, there will be a sequence-enriched region near the TSS region.

Density Map This subgraph visualizes the enrichment status of the ChIP-Seq data in the previous subgraph. You can visually observe a peak appearing near the TSS region. The shape of the peak may reflect the type of sequences. Typically, sharp peaks correspond to TFs(Transcription Factors), while broad peaks correspond to DNA methylation.

Venn Plot Due to the distribution of known TSS regions, a single peak may have multiple potential attributes. For example, when two TSS regions are in close proximity (2~3kb), a peak between them may have both Intergenic and promoter attributes. Therefore, peaks with multiple similar attributes were clustered and a Venn plot was generated to describe the features of all ChIP-Seq data.

Pie Plot This subgraph classified all peaks based on their locations, resulting in a clear pie chart reflecting the proportion of each peak type. You may want to pay particular attention to the proportion of promoter, which may reflect the proportion of TF sequences in all ChIP-Seq.

TF Binding Loci Distribution Map This subgraph depicts the distribution (predicted results) of all possible TF binding sites near the TSS region, and distinguishes the relative distance between each binding site and the TSS region by color.

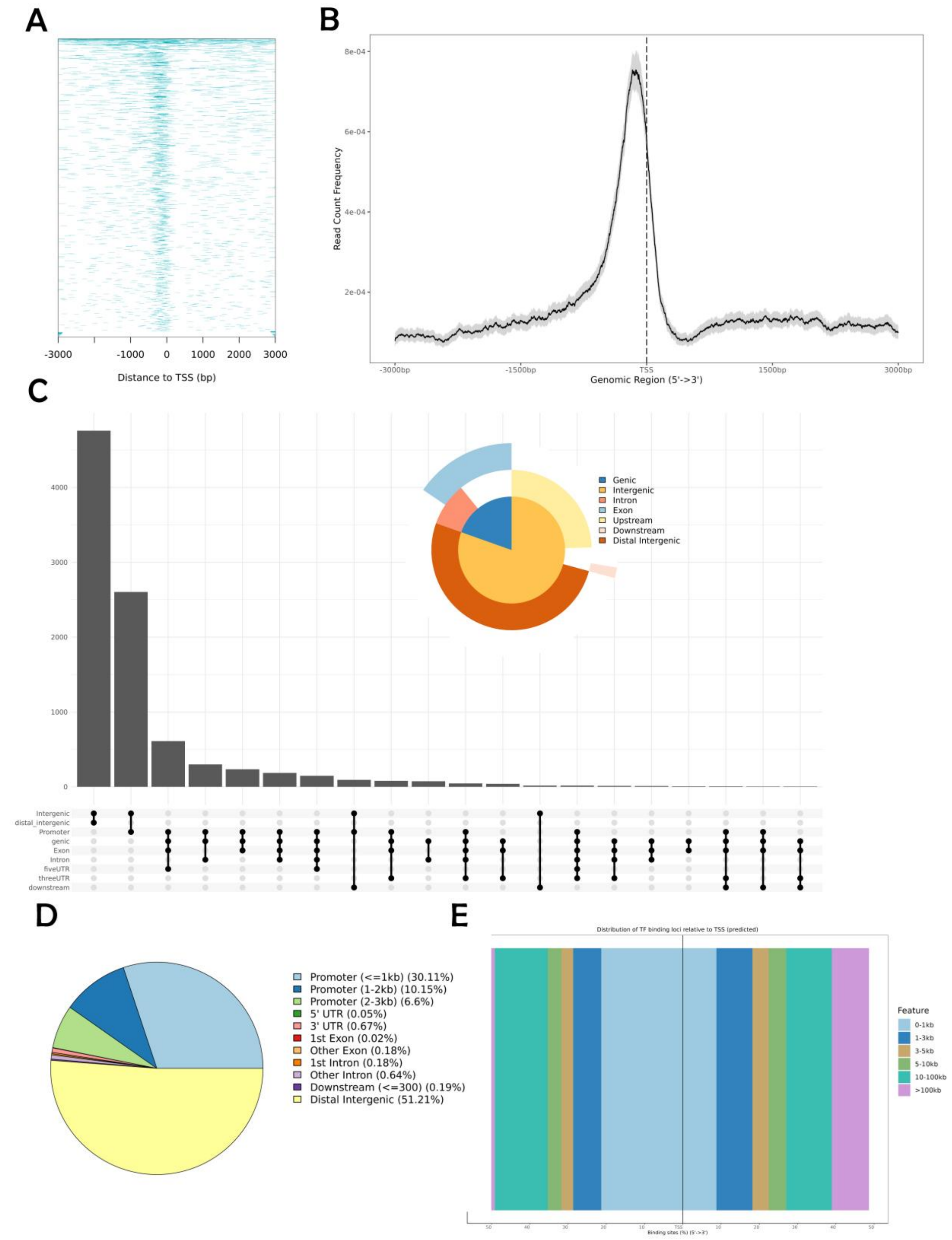


Figure 6.3.2 results of ChIP-Seq Annotation Analysis A. peak distribution heatmap B. peak distribution density map(interactive) C. proportion of peaks within certain genomic regions(venn plot) D. proportion of peaks within certain genomic regions(pie plot) E. distribution map of TF binding loci relative to TSS(interactive)

Chapter 7

7 Data Warehouse

7.1 The HEMU user guide

To provide users with a concise understanding of the various **Toolkits** offered on this website, we have prepared a user guide. This guide outlines the overall framework of the website and provides instructions for using each module. You can view it online or download the PDF. As the website's features evolve, the content of the user guide will be adjusted accordingly.

7.2 Resources

All data used for display and visualization has been categorized into three sections: Genomic Resources, Transcriptomic Resources, and Epigenomic Resources. You can refer to these sections at any time to explore the data that interests you.