



HEMU

database

Contents

Chapter 1

Introduction

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).



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.

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.

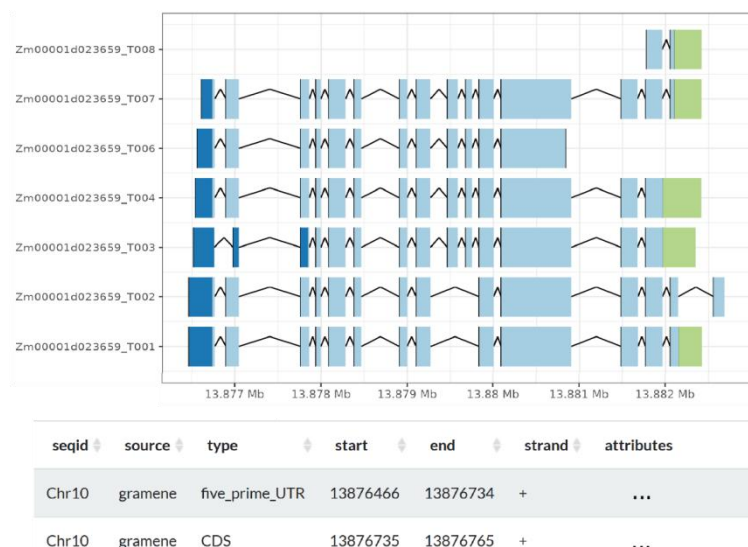


Figure 2.1.2 results of Gene Structure Visualization and details of each gene element

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.

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

Chapter 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.

The screenshot shows the 'Gene Expression Database: Main Search' interface. It features a text input field for Gene IDs containing 'CI036228;CI020599', a dropdown menu for species selection currently showing 'Coix lacryma-jobi var.lacryma-jobi (GCA_009763385.1)', and a format selection section with radio buttons for 'in FPKM' and 'in TPM' (the latter is selected). Below these are two buttons: 'Show Example' and 'Search Expression Profile'. Red arrows point from labels to these specific elements: 'Gene ID Input Box' to the text field, 'Species Select' to the dropdown, and 'Format Select' to the radio buttons.

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.



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 ";".

Figure 3.2.1 original interface of Raw Sequence Acquisition(example showed)

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.

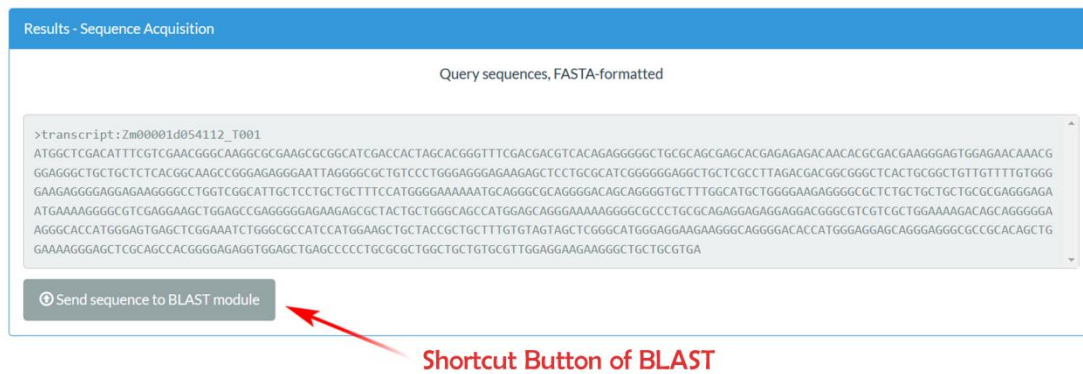


Figure 3.2.2 results interface of Raw Sequence Acquisition

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.

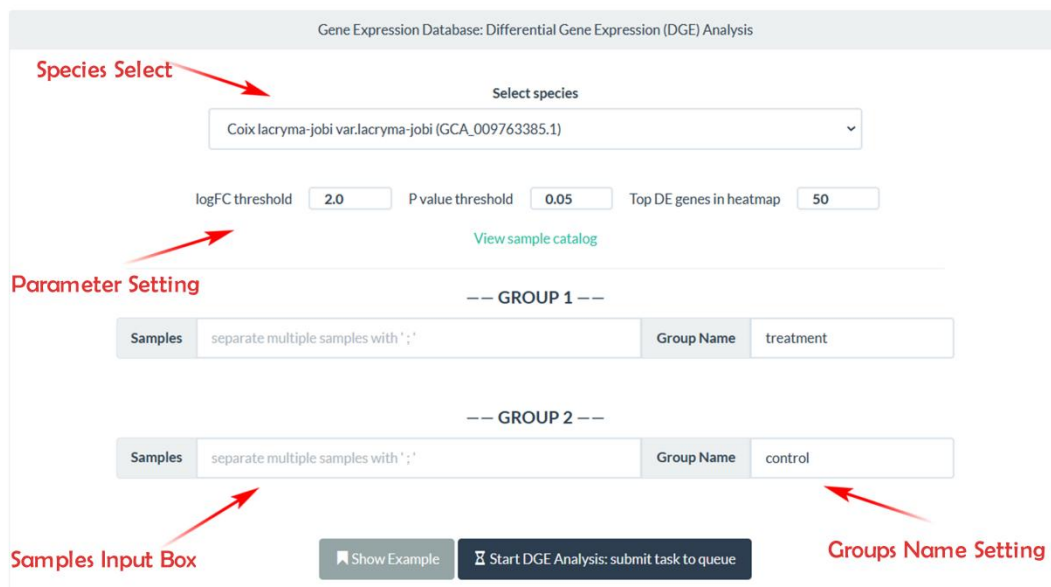


Figure 3.3.1 the interface of Differential Gene Expression Analysis

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 In 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 This 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

Species to enrich

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

Enrich GO Enrich KEGG

GO/KEGG Select

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

Show Example Enrich: submit task to queue

Gene List

Submit Button

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.

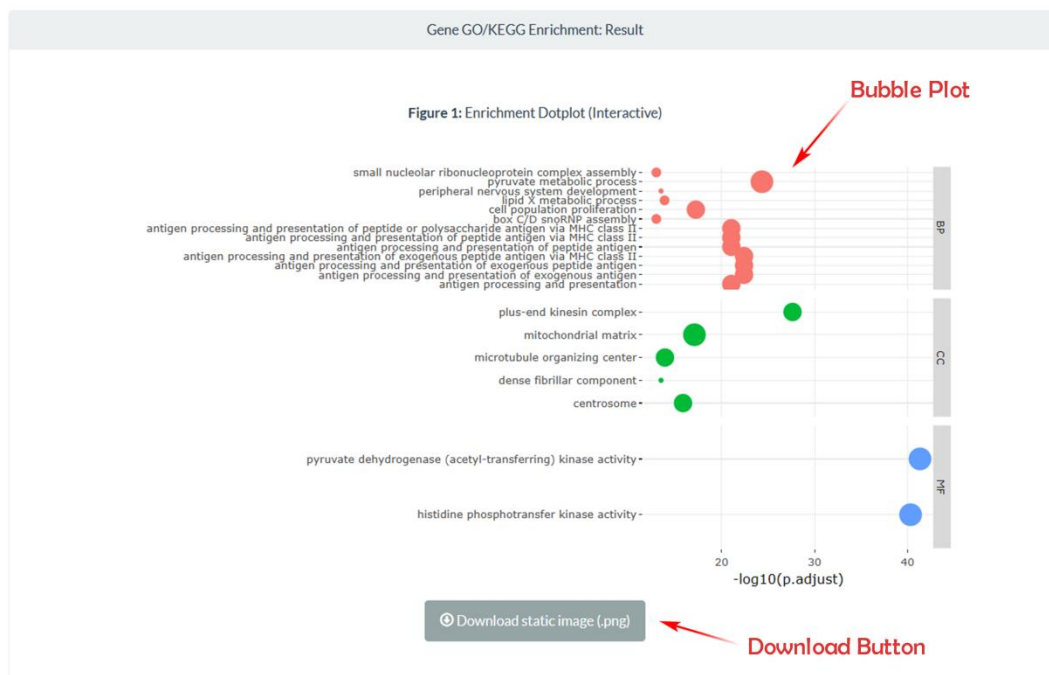


Figure 3.4.2 bubble plot of GO/KEGG Enrichment result

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)

3.6 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.

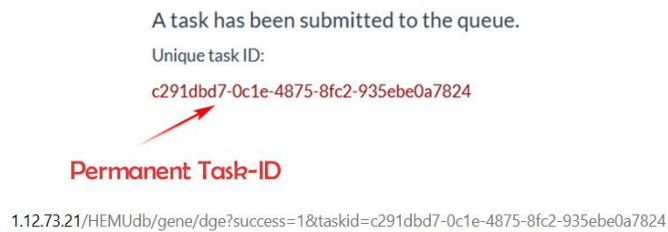


Figure 3.6.1 an example of Task-ID and its place on URL

Chapter 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.

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.

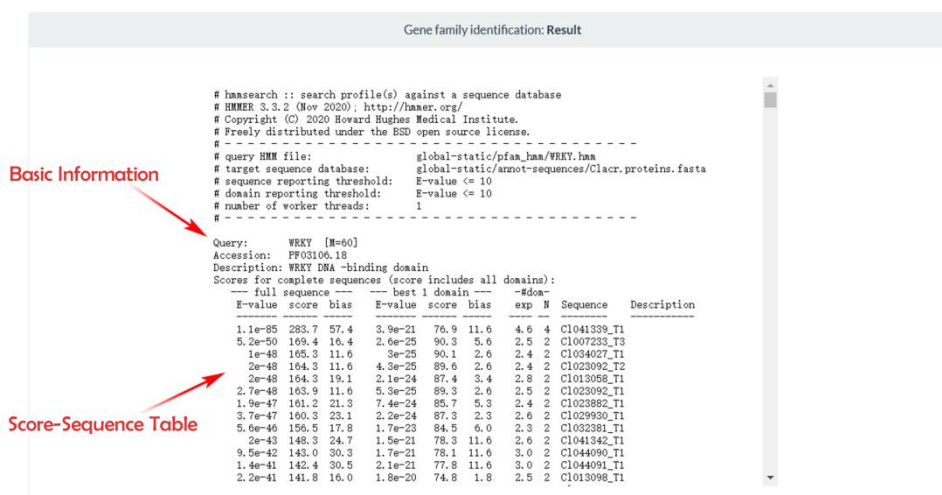


Figure 4.1.2 A example of result based on HMM

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.

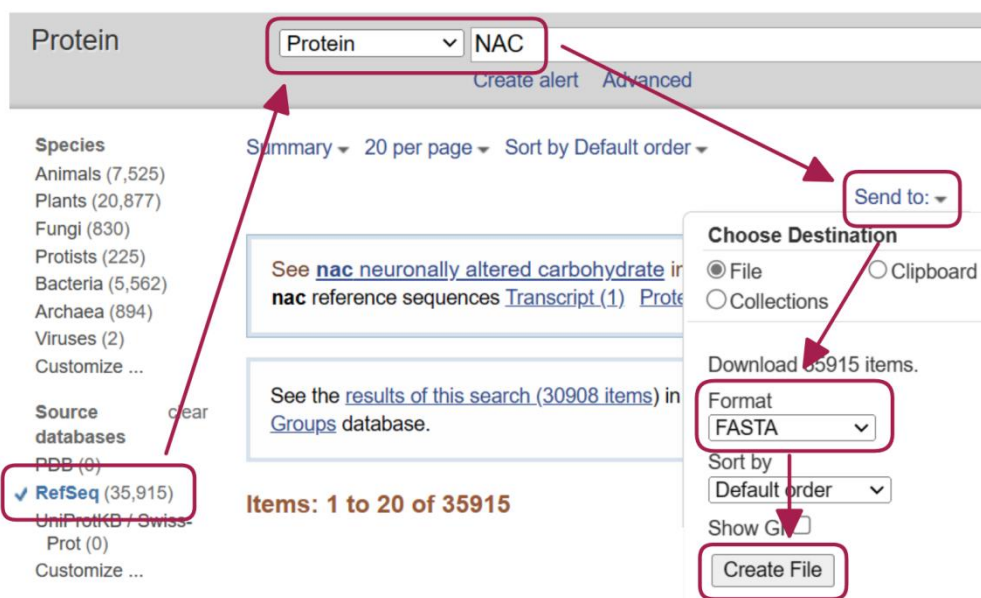


Figure 4.1.3 guide for downloading RefSeq from the NCBI database

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.

Figure 4.2.1 the interface of Phylogenetic 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*.)

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 1: Gene expression level heatmap and sample cluster dendrogram (in TPM)



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

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 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., Sanniguel, 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](#)

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](#)

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 TE ID Query by sample accession ID

Query by SAMPLE 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](#)

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.



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 preset Gene ID/sequence region range will be filled in the corresponding box automatically.

TE analysis toolkit: TE Insertion Location Search

Species Select

Select species dataset: ▼

Select flanking region length for search (bp, ±): ▼

Query flanking TE insertions by **gene ID** [Show Example](#)

Gene ID:

-- or --

Query flanking TE insertions by **sequence region range** [Show Example](#)

Sequence & Range:

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

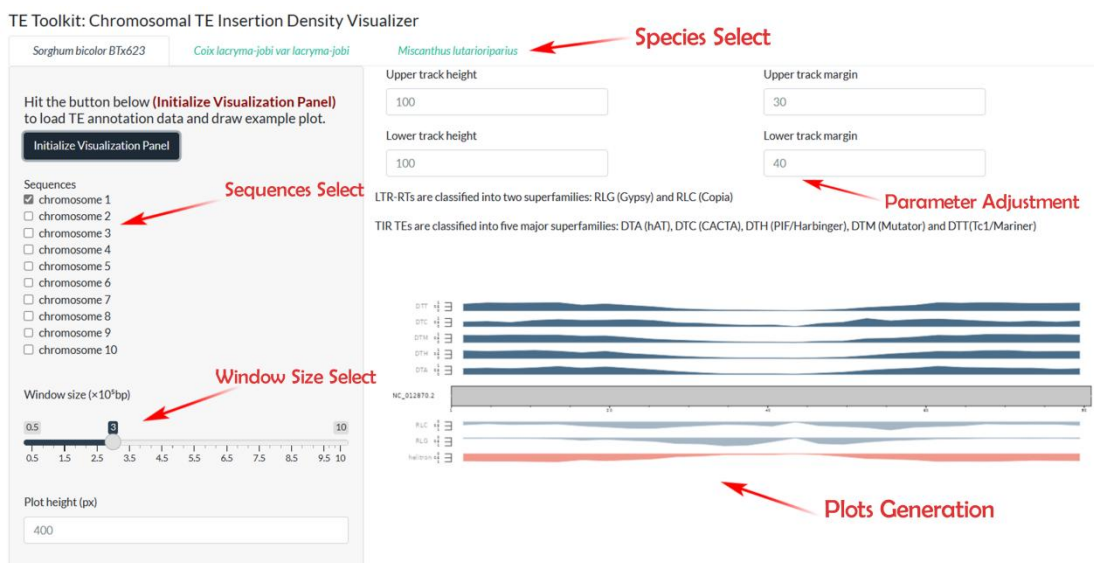


Figure 5.3.1 interface of TE Insertion Density Visualizer(interactive)

About the results, a series of TE insertion density example plots will be generated. Defaulted window size is 3×10^5 bp.

Chapter 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 preset Gene ID/sequence region range will be filled in the corresponding box automatically.

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 TFBS Search

6.3 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 preset sample accession ID will be filled in the boxes automatically.

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.

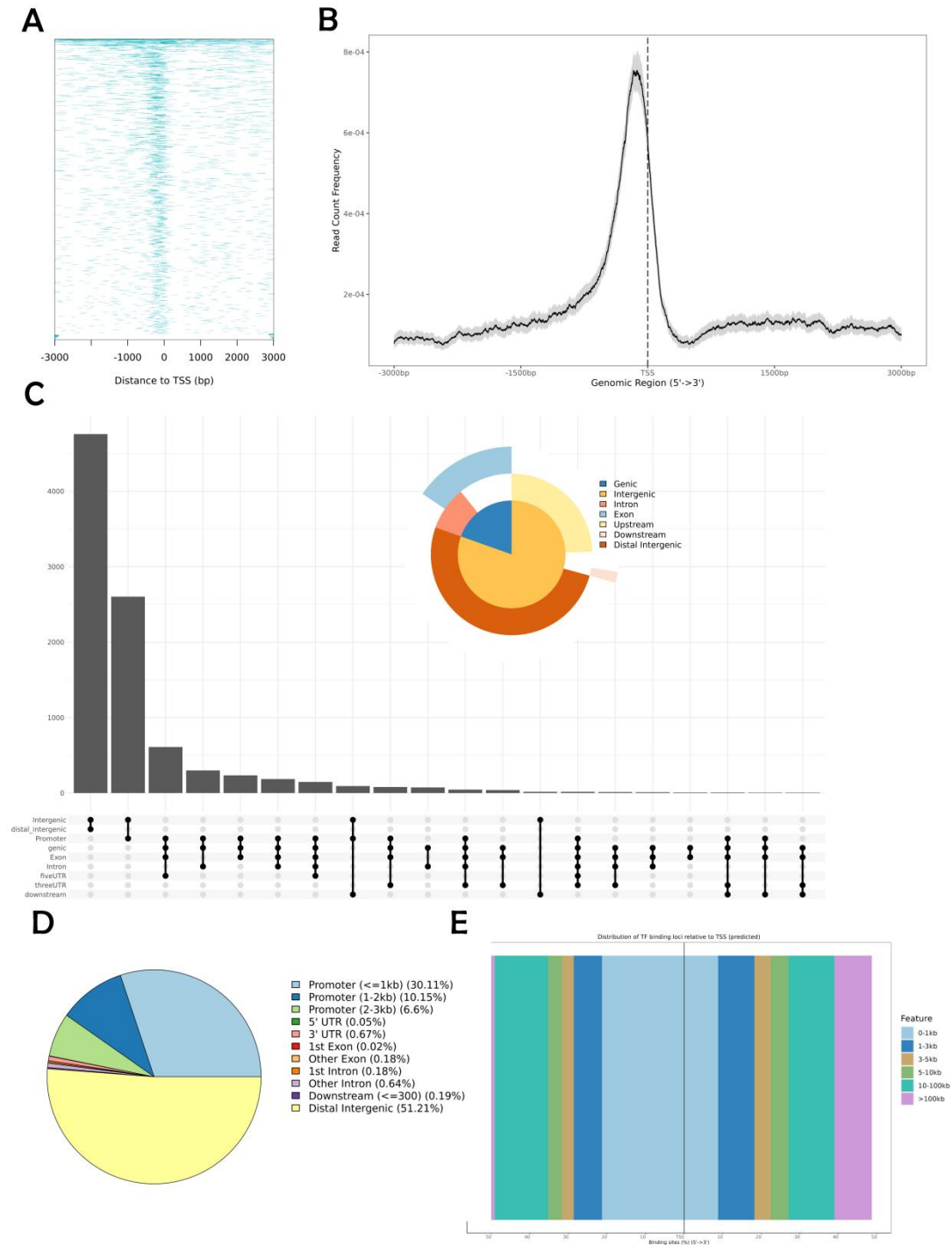


Figure 6.3.2 results of ChIP-Seq Annotation Analysis A. peak distribution heatmap B. peak distribution density map(interactive) C. proportion of peaks within certain genomic regions(venn plot) D. proportion of peaks within certain genomic regions(pie plot) E. distribution map of TF binding loci relative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

6.4 ATAC-seq Peak Annotation Analysis

Chapter 7

Miscellaneous Function

7.1 JBrowse

7.2 BLAST