RSVdb User manual

A comprehensive database of in vivo mRNA structures

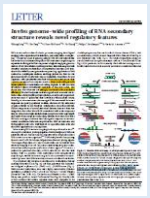
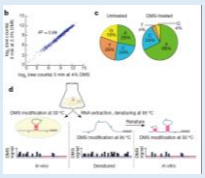
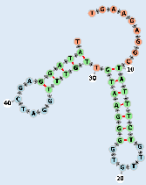
Contents

1.Overview

1.1 The data contained in the RSVdb

RSVdb is a comprehensive database of *in vivo* RNA structure, which contains almost all the genome-wide RNA structure data *in vivo* using DMS (Dimethyl Sulfate, DMS) in the experimental step, one of the most popular agent in RNA structure probing. The method probing RNA structure data involved DMS was first introduced by Rouskin (Rouskin et al. 2014) and Yiliang.Ding (Ding et al.2014) in 2014. RSVdb can be accessed via (**https://taolab.nwsuaf.edu.cn/rsvdb/**).

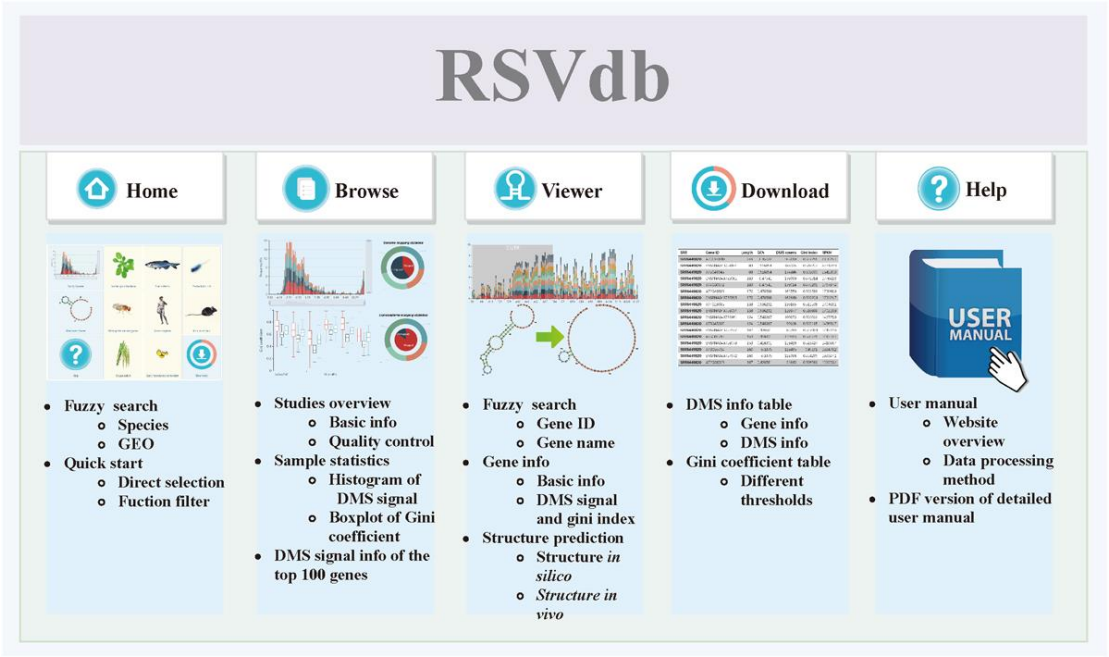
Our database collected all studies since 2014 that used DMS technology to label the structure of mRNA *in vivo*. A total of 8 species, 10 studies, 178 samples, and 622,429 RNAs (with valid DMS data) were collected.

Species	Studies	Samples	RNAs(with valid DMS data)
			
8	**10**	**178**	**622,429**

In addition to the DMS method, there are other methods such as SHAPE-related method or CMCT to mark the mRNA structure *in vivo*, which will be continuously integrated and updated into the RSV database.

1.2 The front-end and main pages overview

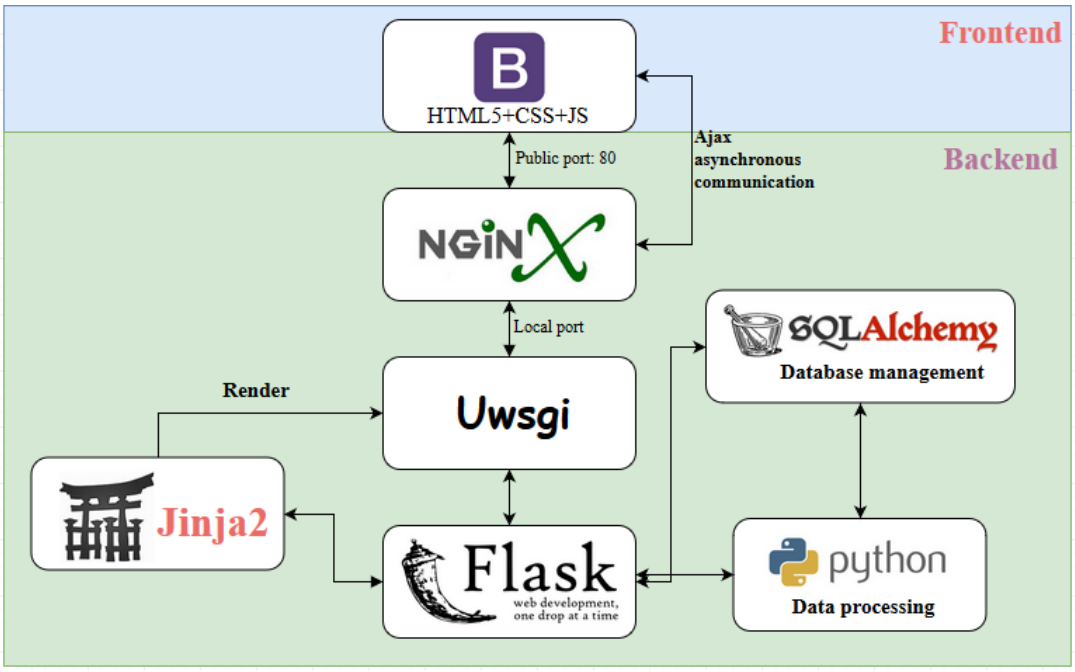
There are five sections of this website, **Home**, **Browse**, **Viewer**, **Download,** and **Help**. The navigation bar on the top gives user entry to each section.



- **Home:** Search for species or research of interest by fuzzy search or enter the other pages of RSVdb by‘Quick Start’.
- **Browse:** Current information about the studies and samples on the mRNA structure *in vivo*, including basic information, sequencing quality control, mapping statistic, RPKM distribution, Gini coefficient distribution, display of RPKM top 100 genes information, etc.
- **Viewer:** Search for RNA of interest, predict RNA structure and visualize the *in silico* and in experiment RNA structure. RNA structure was predicted by using minimum free energy and in experiment RNA structure modification under normalized DMS signal of different experimental conditions.
- **Download:** Download the TSV version of all the data provided by RSVdb.
- **Help:** RSVdb user manual, detailed website introduction, our contact information, and acknowledgments.

The front-end of the website used HTML5+CSS+JS layout. In order to give users a better experience, all diagrams are interactive figures (that is, drag, zoom, image download, mouse events or other functions). The front-end and back-end data communication used AJAX asynchronous communication, so it is smoother to update data and charts.

1.3 The back-end overview



The backend of RSVdb

The back-end of the website is mainly built by the flask-web framework, including SQLalchemy for database management, Jinja2 for template rendering, and Python scripts for data processing and calling local software (eg: 'RNAstructure'). Flask work with uWSGI and NGINX to handle the requests.

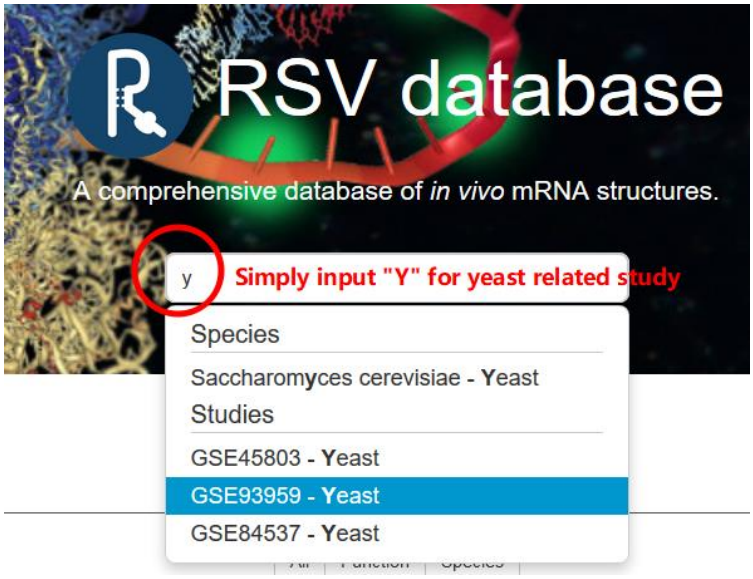
2. Manual of each section

2.2 Home

The home page includes two functions: **fuzzy search** and **quick start**. Users can easily access all pages on the website from the ‘Home page’.

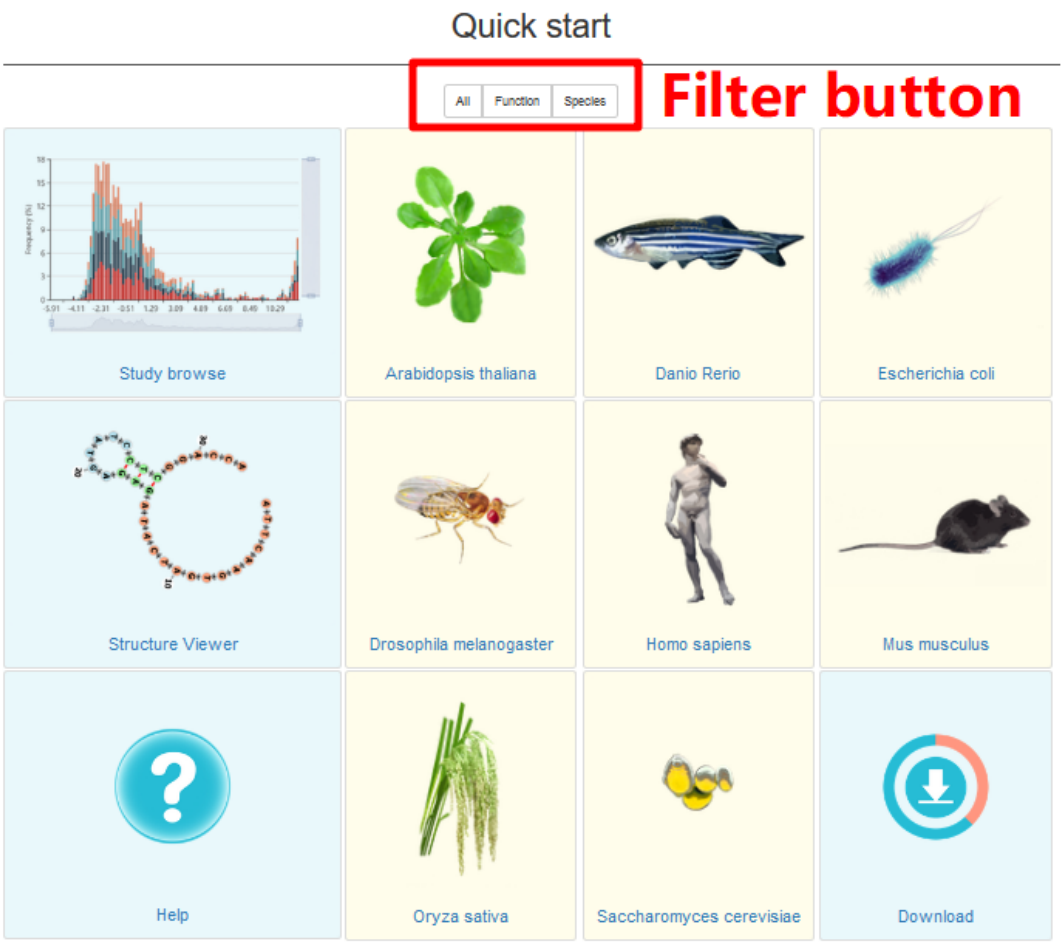
2.2.1 Fuzzy search

You can input species, GEO number or SRP number in the search box. Fuzzy search makes it easy to search for **species** and **research** by simply typing in a few letters.



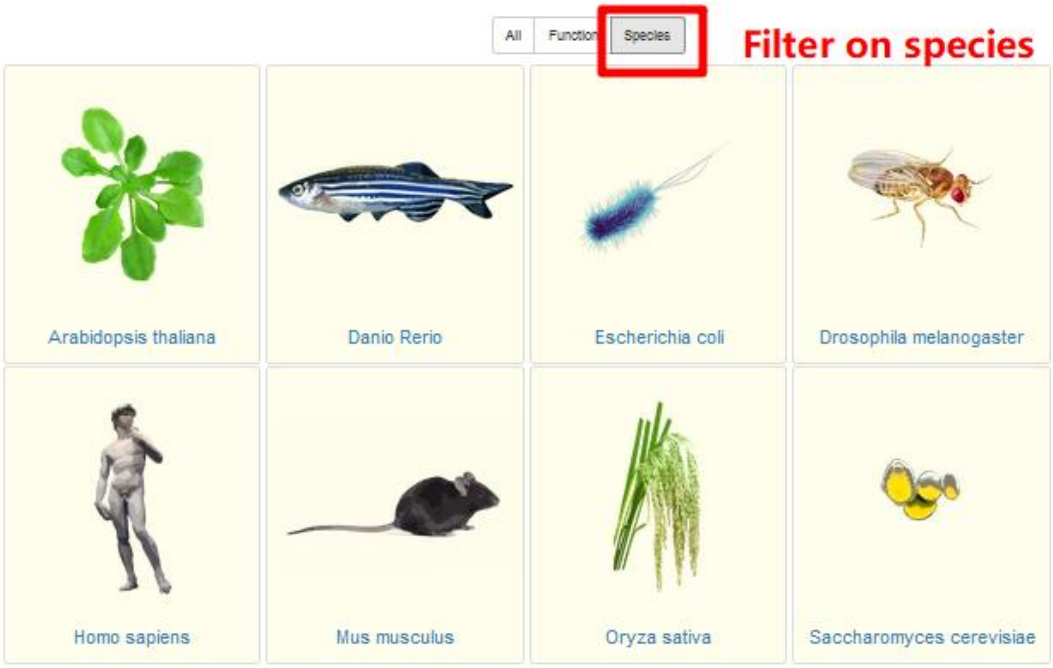
By selecting the right species or research, you can jump straight to the Browse page and display the relevant information.2.2.2: Quick start

In order to make it easier for you to use RSVdb, we designed a set of pictures typesetting in the 'Quick start' section. You can click one of the species icons (Yellow) provided and enter the browse page of the species. Or click one of the function icons (Blue), and you can quickly enter the corresponding section.



In order to prevent future species or function increase, we also designed the screening button. Click on the corresponding button to filter the corresponding icons of "website function" or "species".

Filter the function pages of this site.



Filter the species this site contains.

2.3 Browse

The Browse page allows you to view information about a study, including basic information, sequencing quality control, mapping statistic, RPKM distribution, Gini coefficient distribution, display of RPKM top 100 genes information, etc.

The navigation of this page on the left includes species and studies. You can choose the species you are interested in and research accordingly. Then you will see the main information about the study and basic statistical analysis of the *in vivo* RNA structure data.

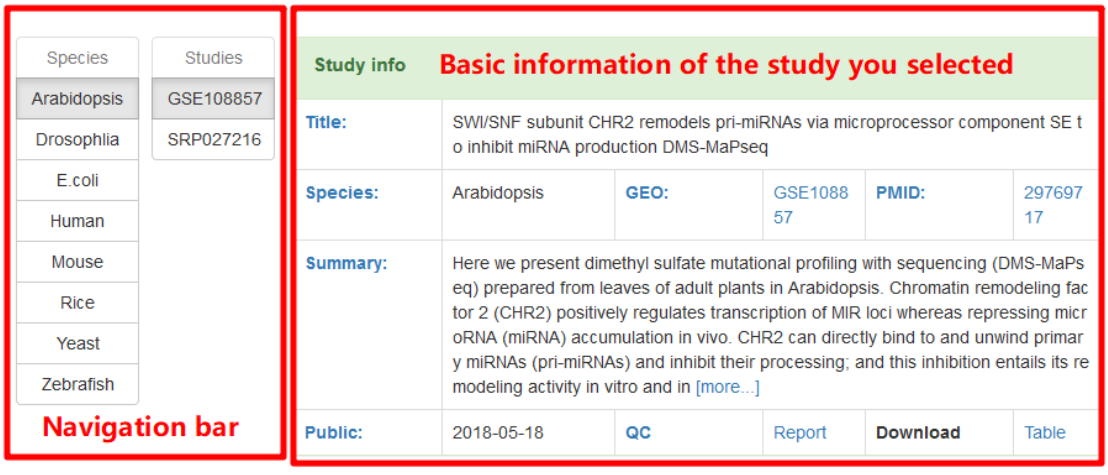
2.3.1 Basic information

We present the basic information about the study from three dimensions: study, sample and SRR files. The relationship between them is that one study (research) contains several samples (experimental treatment), and one sample contains several SRR (sequencing file, SRRxxxxx.sra). It is worth saying that 'sample' actually refers to the experimental treatment, such as temperature gradient treatment, reagent treatment with different concentrations, etc. This is to allow users to compare the differences between different experimental treatment groups.

2.3.1.1 Basic information of study

2.3.1.1.1 Study info

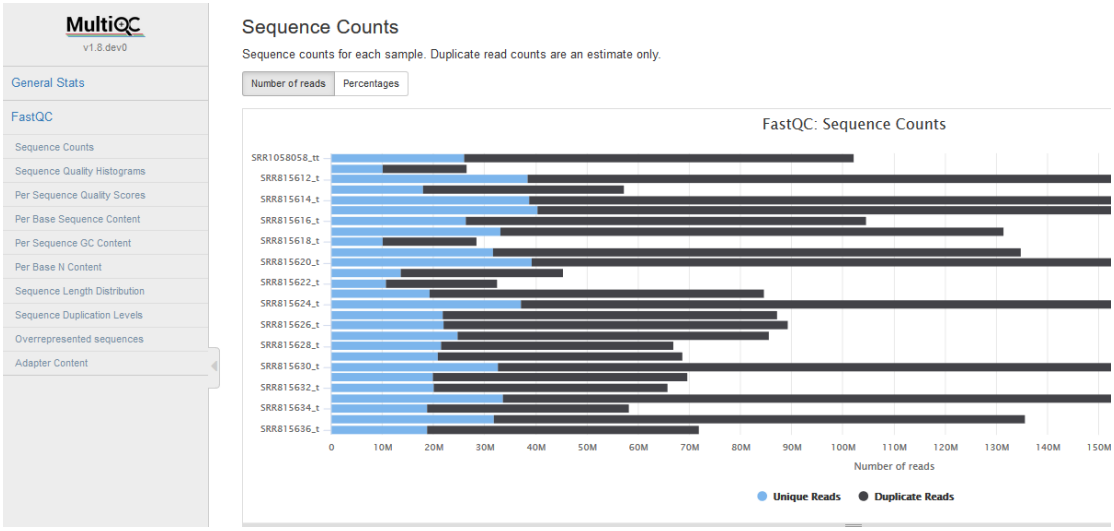
The panel, **'Study info'** is the basic information of the study you selected.



- **Navigation bar:** On the left is the species navigation, which contains eight species. On the right is the research navigation, which contains all the research on this species.
- **Study info:** This section contains key search information for the study, including study title, in which species, GEO number, PMID number, a summary of the study, the publication date

of the study, QC hyperlinks, and downloads available for the study.

- **Title: The title of data source study**
- **Species: The species involved in the data.**
- **GEO: Accession number of the Gene Expression Omnibus(GEO) database**
- **PMID:** PubMed Unique Identifier (PMID).
- **Summary:** Summary of this study experiment design, retrieved from GEO database.
- **Public:** Public time of the data or study
- **QC:**

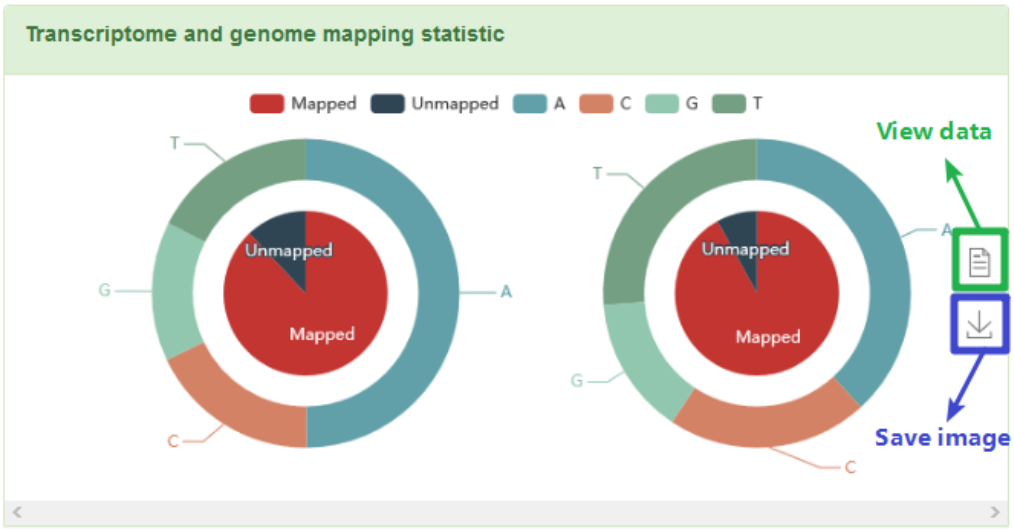


The QC hyperlink contains a comprehensive comparison of data quality control made by MultiQC under different experimental conditions in this study, which was obtained by FastQC.

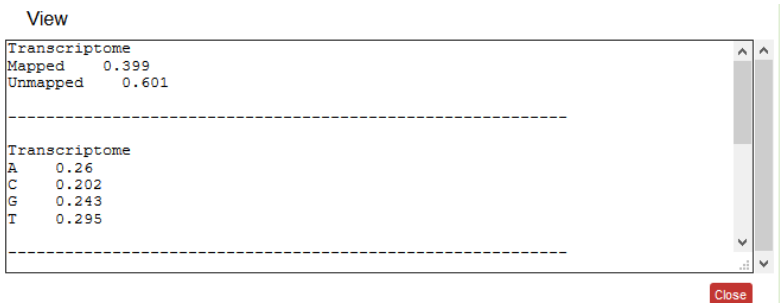
- **Ref:** Reference genome for Mapping.

2.3.1.1.2 Transcriptome and genome mapping statistic

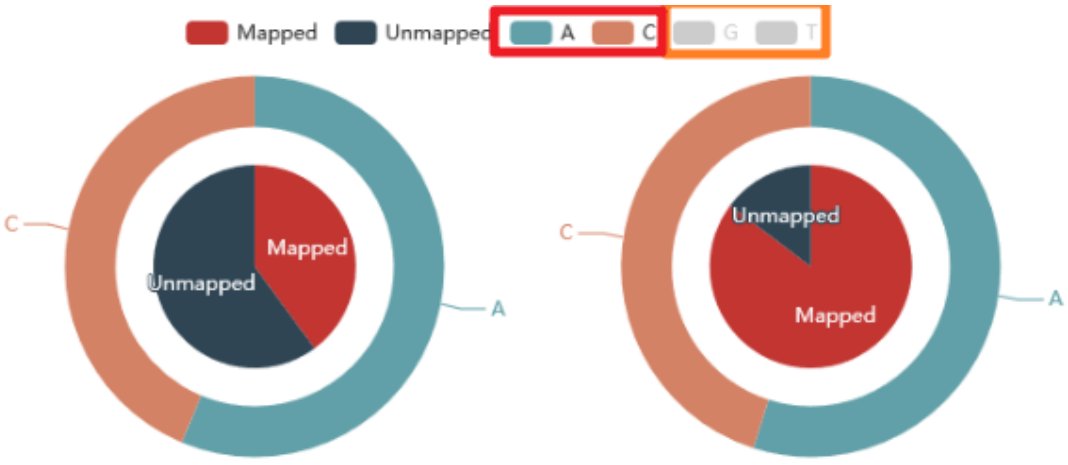
The panel, **'Transcriptome and genome mapping statistic'** compared the DMS data of this study in mapping transcriptome and genome. This chart supports data customization, data browsing, and image saving.



The pie map includes the proportions of mapped and unmapped reads. The differences in the total mapped reads were compared. The circular diagram includes the ratio of DMS signal markers to four bases in A, G, C and T, and the difference of DMS signals between transcriptome (left) and genomic (right) in these four bases are obtained by comparing the circular diagram.



View data



Data customization

2.3.1.2 Basic information of the sample

2.3.1.2.1 Sample contains

The panel, '**Sample contains**' includes all sample information for the study.

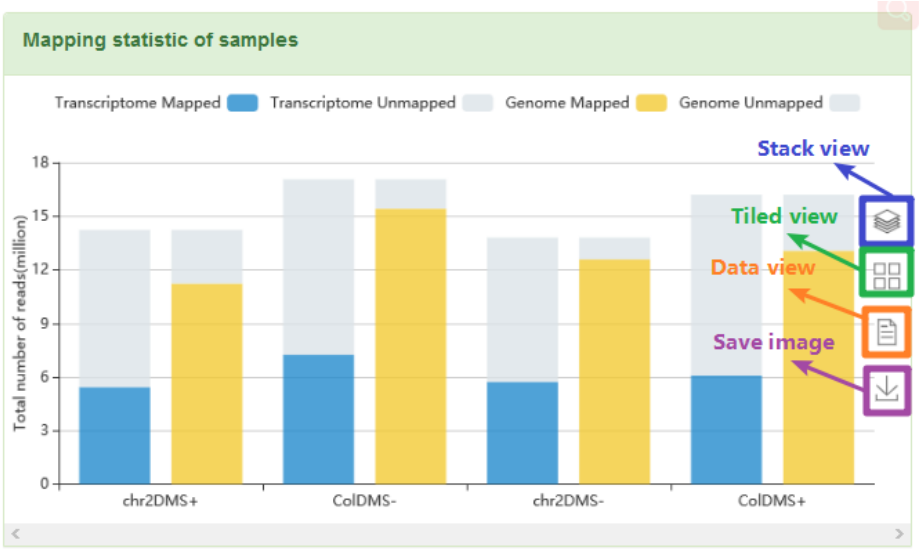
Sample contains

Label	Sample	SRR list
chr2DMS+	chr2 DMS (+)	SRR6449824,SRR6449822
ColDMS-	Col-0 DMS (-)	SRR6449819
chr2DMS-	chr2 DMS (-)	SRR6449820
ColDMS+	Col DMS (+)	SRR6449821,SRR6449823

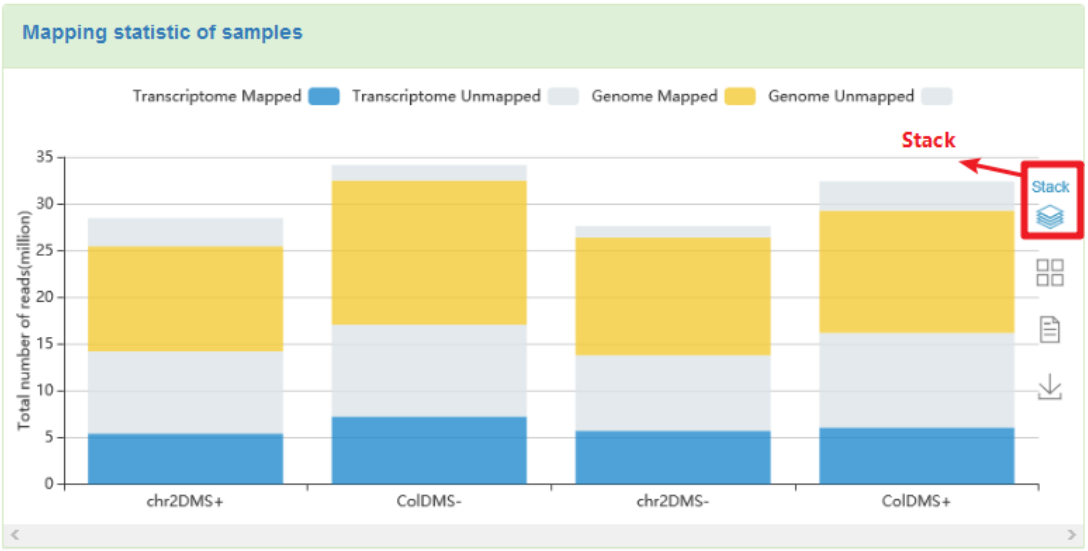
- **Label:** Short names of each experiment sample. The original samples name was replaced by the abbreviation.
- **Sample:** Complete sample name, which is equal to the GEO database.
- **SRR list:** All accession numbers of SRR database for each sample.

2.3.1.2.2 Mapping statistic of samples

The panel, "**Mapping statistic of samples**" contains statistics of each sample mapping to transcript and genome DMS signal reads. The ordinate is the mapped DMS signal reads in per million. This chart supports data customization, stack and tiled display, data browsing and image saving.



Stack display

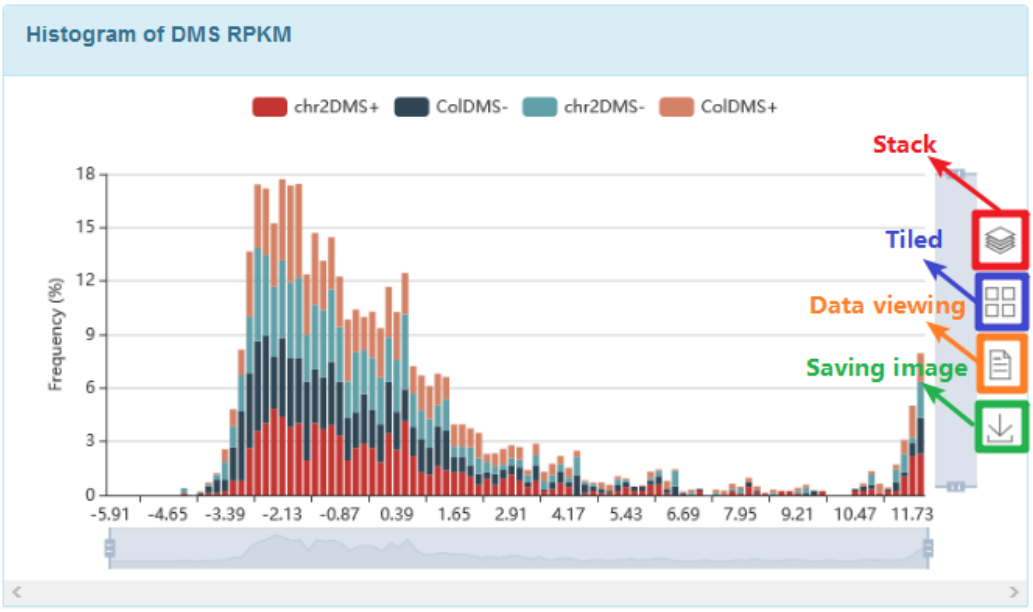


Tiled display

Click ‘Stack’ or ‘Tiled’ icon, you can change the histogram style to Stack and tiled, respectively.

2.3.1.2.3 Histogram of DMS RPKM

The panel, ‘**Histogram of DMS RPKM**’, contains the RPKM frequency histogram calculated from the DMS signal value of each transcript for each sample. This chart supports data customization, stack and tiled display, data browsing and image saving.

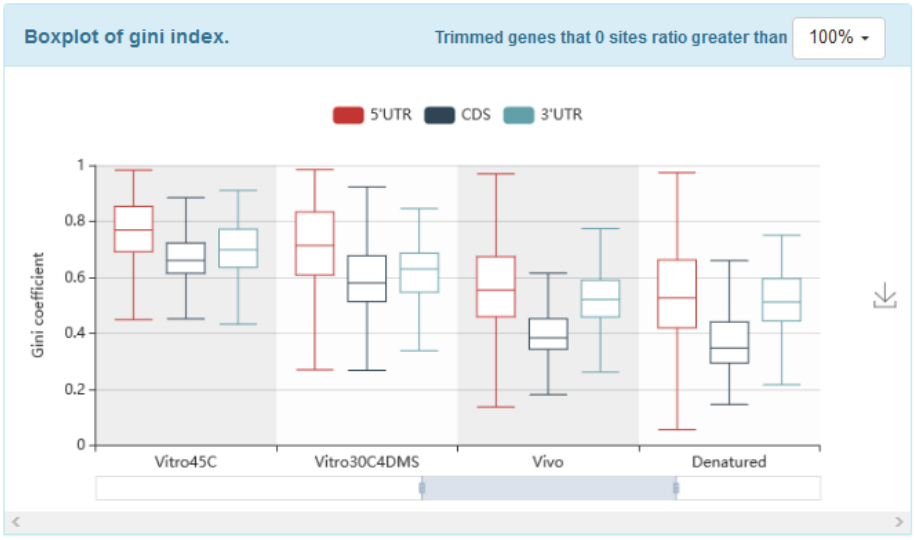


Clicking the sample color label to gray does not display the sample data. RPKM is defined as dividing reads mapped to transcripts by all mapped reads of the transcriptome (in millions) and transcript lengths (in kilobytes).

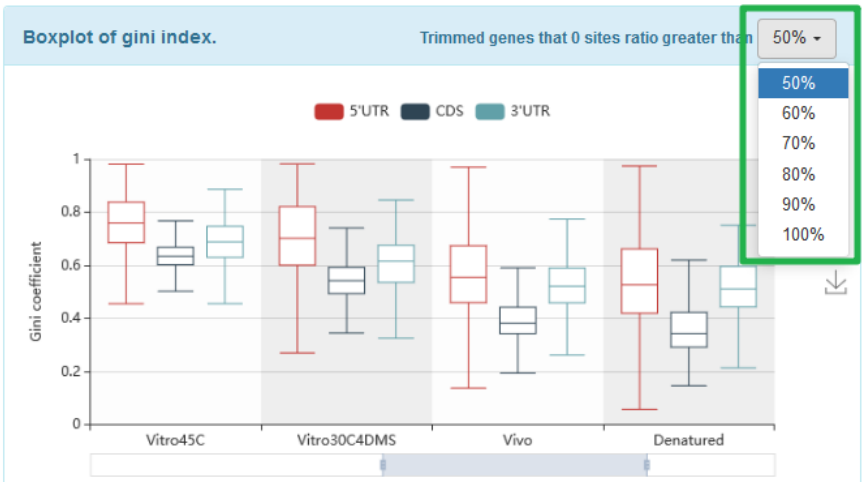
2.3.1.2.4 Boxplot of Gini coefficient

The panel, '**Boxplot of Gini coefficient**', contains the Gini coefficient values of different regions (5'UTR, CDS, 3'UTR) of the transcript in different samples. This chart supports custom thresholds and image saving.

Gini coefficient could be a measure of the strength of mRNA structure (Rouskin et al. 2014). We provide boxplot of the Gini coefficient of different regions in different samples of one study.



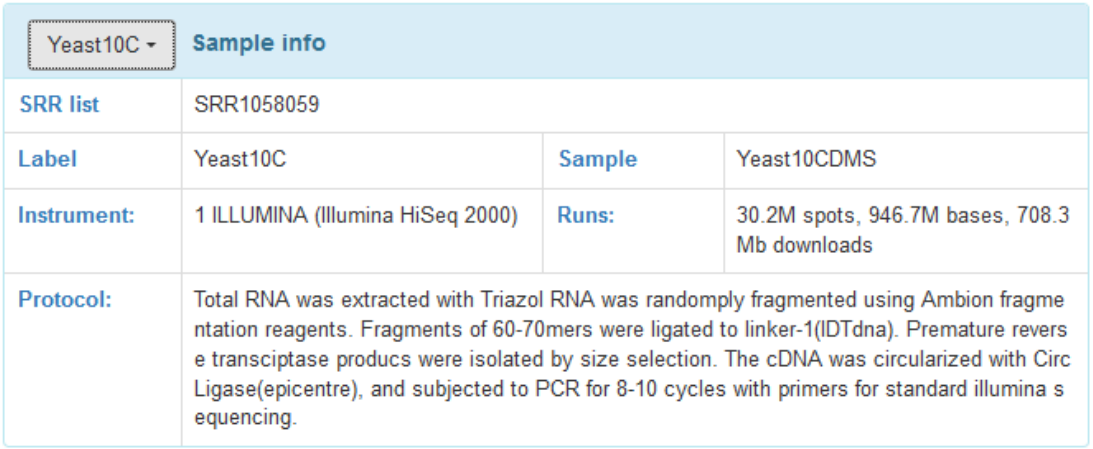
The Gini coefficient of each transcript is calculated by sliding window, with a size of 50 A and C bases (G and T bases are removed) and a step size of 50nt.



The drop-down menu on the right can choose to retain the Gini coefficient calculated by different thresholds at site 0 ratio (DMS signal's coverage).

2.3.1.2.5 Sample info

The panel, '**Sample info**', contains the basic information about the sample, including the sample name, data volume, sequencing, and the experimental protocol.

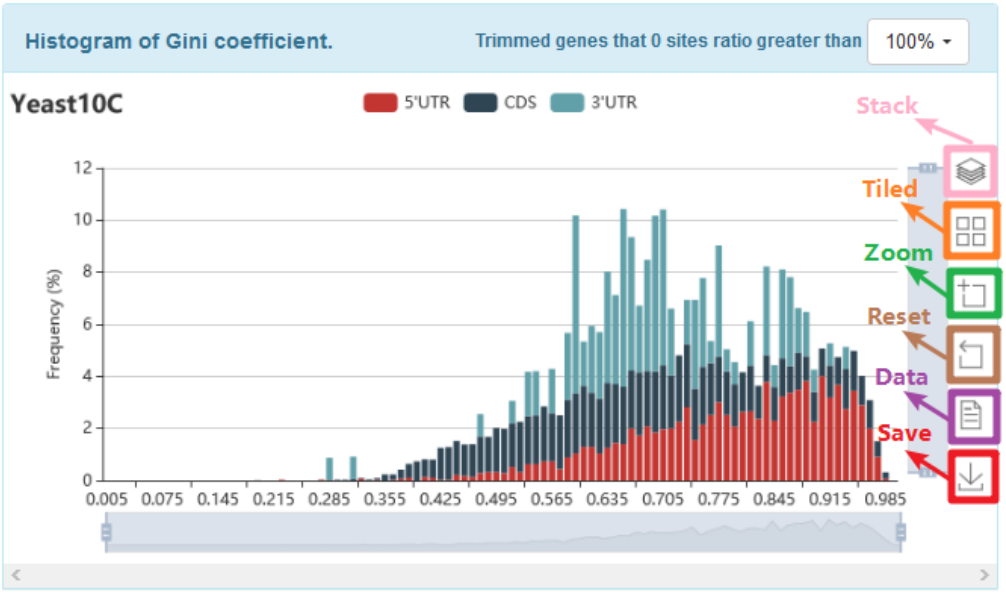


- **SRR list: The list containing accession number of Sequence Read Archive database**
- **Label: Short name of each experiment samples. The original samples name was replaced by the abbreviation.**
- **Sample: The complete sample name used in the study, obtained from GEO database**
- **Instrument: Sequencer used**
- **Runs:Statistics of reads in the sample**

- **Protocol:Construction protocol**

2.3.1.2.6 Histogram of Gini coefficient

The panel, '**Histogram of Gini coefficient**', contains frequency histograms of Gini coefficients in different regions (5'UTR, CDS, 3'UTR) of each sample. This chart supports custom thresholds, stack, and tiled display, zoom and resets, data browsing and image saving.

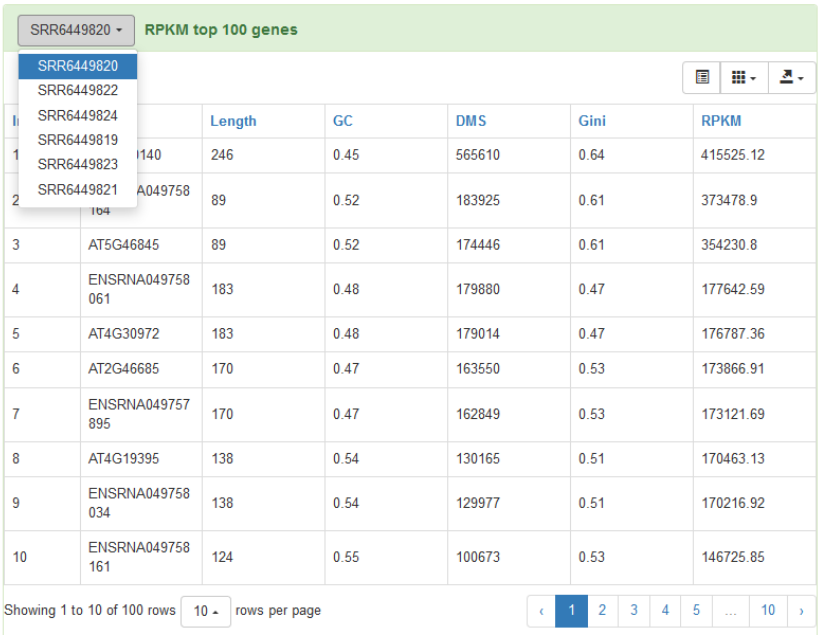


The drop-down menu on the right can choose to retain the Gini coefficient calculated by different thresholds at site 0.

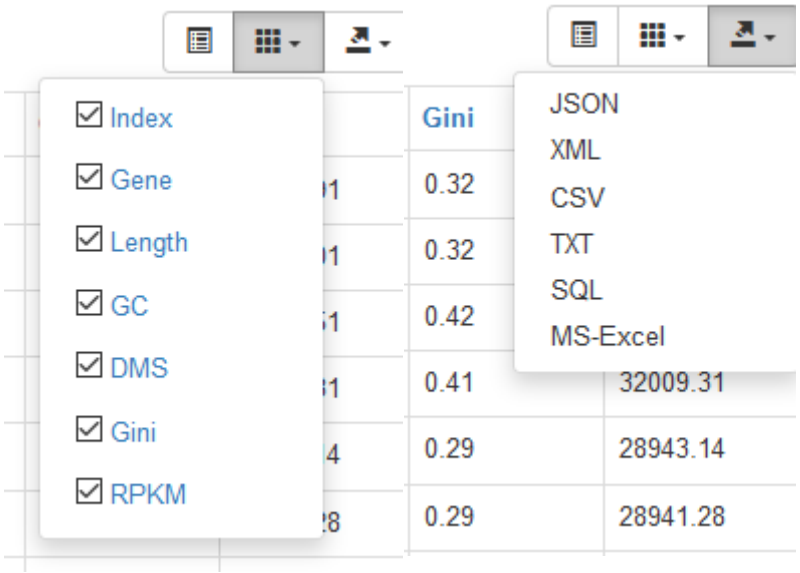
2.3.1.3 Basic information of SRR files

2.3.1.3.1 RPKM top 100 genes

The panel, '**RPKM top 100 genes**', contains the SRR files statistical information about the top 100 genes for sequencing according to the RPKM value calculated by DMS signal. This table supports toggled data display, custom table header, and table export.



- **Index: Number of RPKM top 100 structured genes**
- **Gene: Gene name**
- **Length: Transcript length**
- **GC: GC content of the transcript**
- **DMS: Average DMS signal value(unnormalized) of the transcript**
- **Gini: Average Gini coefficient(The gini coefficient is calculated by sliding window with 50nt window size and 50nt step size) of each transcript**
- **RPKM: Reads Per Kilobase per Million mapped reads(RPKM) of the transcript**



Custom table header and table export.

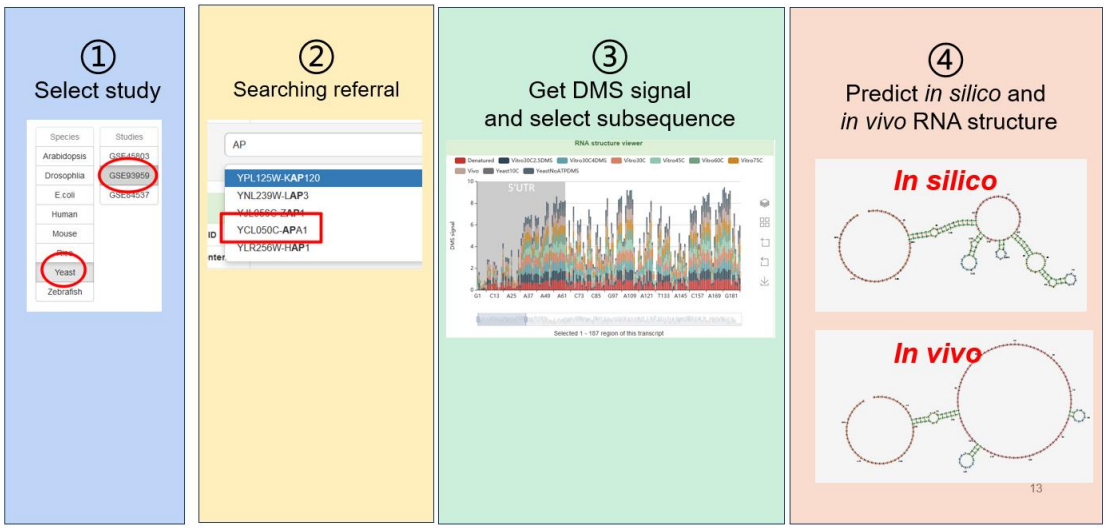
2.4 Viewer

The ‘Viewer page’ is the highlight of the RSVdb database. The ‘Viewer page’ allows you to search the gene you are interested in, displaying DMS signals from the experimental processing of the corresponding research. In addition, users can choose any sub-sequence, combine with their DMS signals to make an in experimental RNA structure prediction and realize visualization.

In this section, we provide the search scope, gene info and RNA structure viewer in the selected study, as showed below.

2.4.1 Easy Using

The RSVdb Viewer is user-friendly. It can be used in four simple steps to get the structure you are interested in (both the analog structure and the DMS signal correction structure).



① The navigation of this page on the left includes species and studies. You can choose which species or research you are interested in and proceed to the next step.

② Search for genes that you are interested in. We support searching referral for more than 660,000 RNAs, which means that if you type in a few letters, RSVdb will recommend genes for selection.

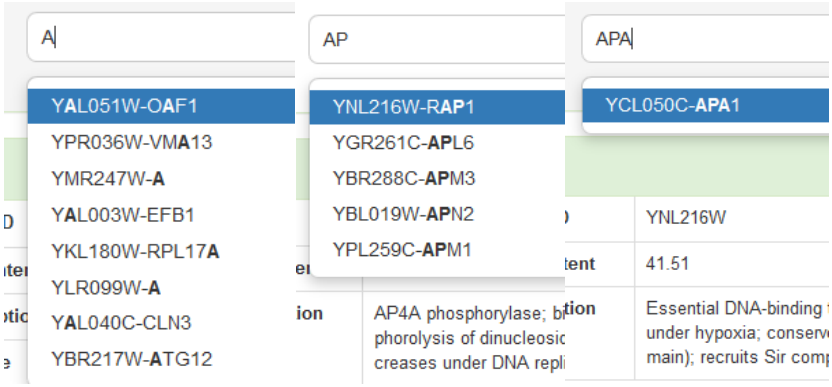
③ After selecting the gene, click or enter to get the gene information and DMS signal. We show the DMS signals in all samples to compare the differences of DMS signals in different experiments.

④ Select a range of the sequence you want to predict and press the button "Generate RNA structure" to predict the structure of the RNA *in silico* and in the experiment.

2.4.2 Search suggestion

Searching referral supports the fuzzy search for gene ID and gene name. Note that this search is limited to the species and studies selected in the first step, i.e. if you select *Saccharomyces cerevisiae*

in the navigation bar on the left, you can only search for yeast-related genes in the search box. For example, if you want to search for the "APA1" gene, just type in "AP "or "APA." Because we processed the data, we screened out the RNA that was lower in DMS signal or lower in quality, so you might not necessarily search for the genes that are involved. Then, you can try other studies of the same species. RSVdb is constantly updated with more data to ensure that as many genes as possible have information about the structure of the mRNA in the body.



Search suggestion of “APA1” gene

2.4.2 Gene info

After selecting the gene, the basic information of its RNA and the distribution of DMS signal can be seen after a short search and waiting. Because it will show the DMS signal of all samples, studies with large samples may search for a longer time (e.g., studies with GSE45803 for 10 samples of yeast pairs may take **more than 10 seconds**).

2.4.2.1 Basic info

The panel, ‘**Gene info**’, contains the gene name, chromosome, gene type, GC content, gene description, DMS signal and Gini coefficient of the corresponding sample of this gene.

Gene Info					
Gene ID	YCL050C	Gene name	APA1	Chromosome	III
GC content	39.34	Gene type	protein_coding		
Description	AP4A phosphorylase; bifunctional diadenosine 5',5'''-P1,P4-tetraphosphate phosphorylase and ADP sulfurylase involved in catabolism of bis(5'-nucleosidyl) tetraphosphates; catalyzes phosphorolysis of dinucleoside oligophosphates, cleaving substrates' alpha/beta-anhydride bond and introducing Pi into the beta-position of the corresponding NDP formed; protein abundance increases under DNA replication stress; APA1 has a paralog, APA2, that arose from the whole genome duplication [Source:SGD;Acc:S000000555]				

Sample	DMS signal	Gini coefficient
Denatured	603.848000000001	0.11
Vitro30C2.5DMS	518.6050000000007	0.19
Vitro30C4DMS	419.47399999999857	0.26
Vitro30C	638.4720000000026	0.13
Vitro45C	534.8019999999997	0.33
Vitro60C	572.249999999999	0.32
Vitro75C	600.335000000001	0.27
Vivo	633.9100000000002	0.11
Yeast10C	350.2619999999997	0.31
YeastNoATPDMS	454.28900000000004	0.21

- **Gene ID:**
- **Gene name:**
- **Chromosome: The gene you searched lies in this chromosome**
- **GC content: GC content of the gene you searched**
- **Gene type: Gene type of the gene you searched**
- **Description: Basic description of the gene you searched**
- **Sample: Sample label list of the gene you searched**
- **DMS signal: Average DMS signal of the gene**
- **Gini coefficient: Average Gini coefficient of the gene(The gini coefficient is calculated by sliding window with 50nt window size and 50nt step size)**

  The gene annotation information is from 'Ensembl' and is empty if the information is missing.

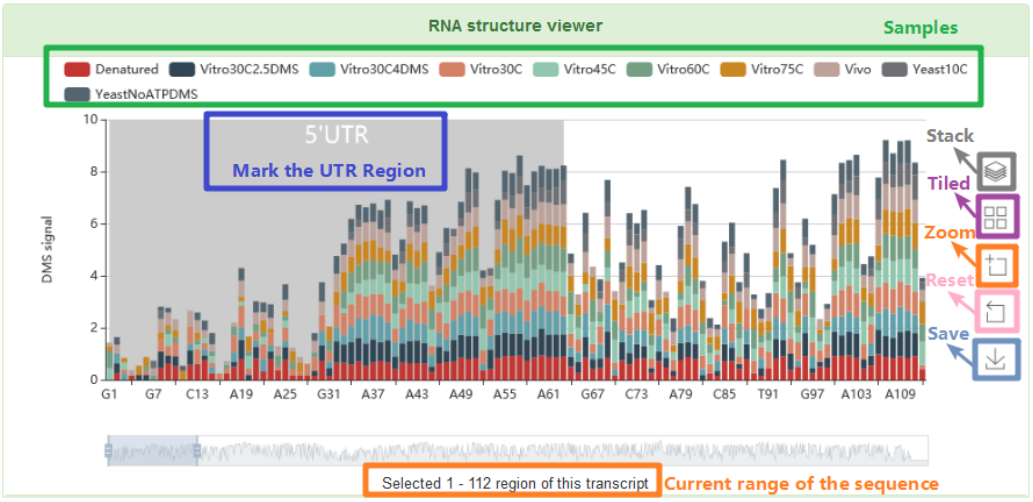
2.4.3 RNA structure viewer

This part is divided into DMS signal distribution and visualization of RNA structure.

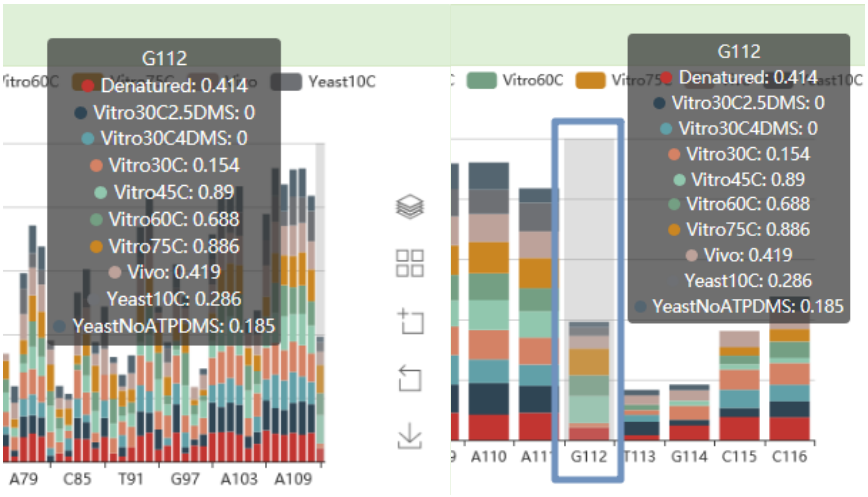
2.4.3.1 DMS signal distribution

In the first part, we provide the DMS distribution at each base of the gene selected of each sample

in the study. This chart supports data customization, stack, and tiled display, zoom and resets and image saving.



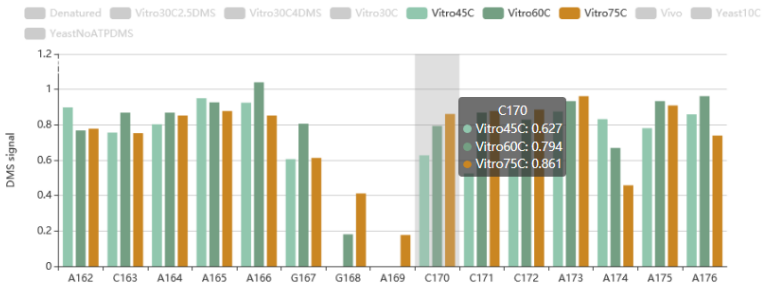
The selected part is the area already shown in the figure. You can hover the mouse over a specific column or zoom to get specific data information. For each base, we mark it with a 'base and position'. For example, 'G112' means the 112th base and 'G'.



Zoom in or mouse hover to get base information

It is worth noting that the DMS reagent can only mark sites **A** and **C**, but at this stage, we are showing signals at all sites. In other words, what are shown is the raw data from the Mapping on the reference genome. DMS signal data are standardized using the method of Ding et al (Ding et al. 2014).

Similarly, the chart supports screening samples, that is, selecting samples of interest and comparing DMS signal differences at different sites.



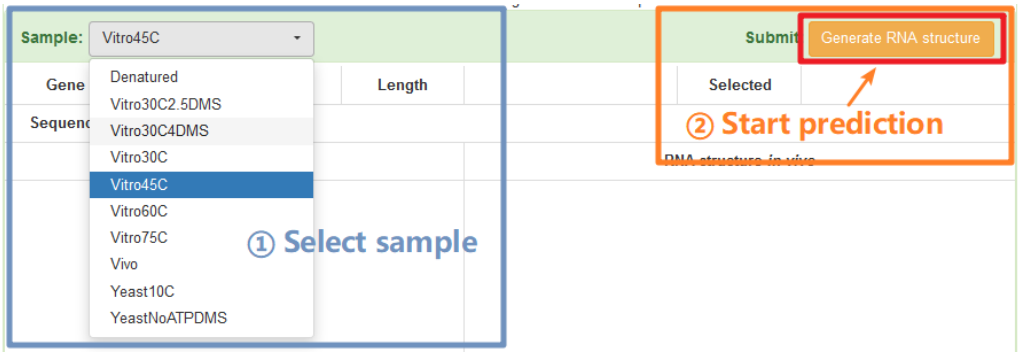
Samples screening.

2.4.3.2 RNA structure viewer

In this part, you can view the interactive RNA structure, both *in silico* and in the experiment (corresponding to the experimental design).

2.4.3.2.1 Data selection and submission

The region shown in DMS signal figure is the sequence to be calculated. To facilitate the selection of specific regions, we recommend using the "Zoom" tool next to the figure to select the regions where you wish to predict RNA structure. Next, select the sample to be predicted from the drop-down menu and click the button to make the prediction.



Note that we do not recommend predicting RNA sequences with a sequence length greater than 500nt, as this reduces the accuracy of the prediction. At the same time, when the forecast length exceeds 500nt, the forecast time increases significantly (tens of seconds, possibly).

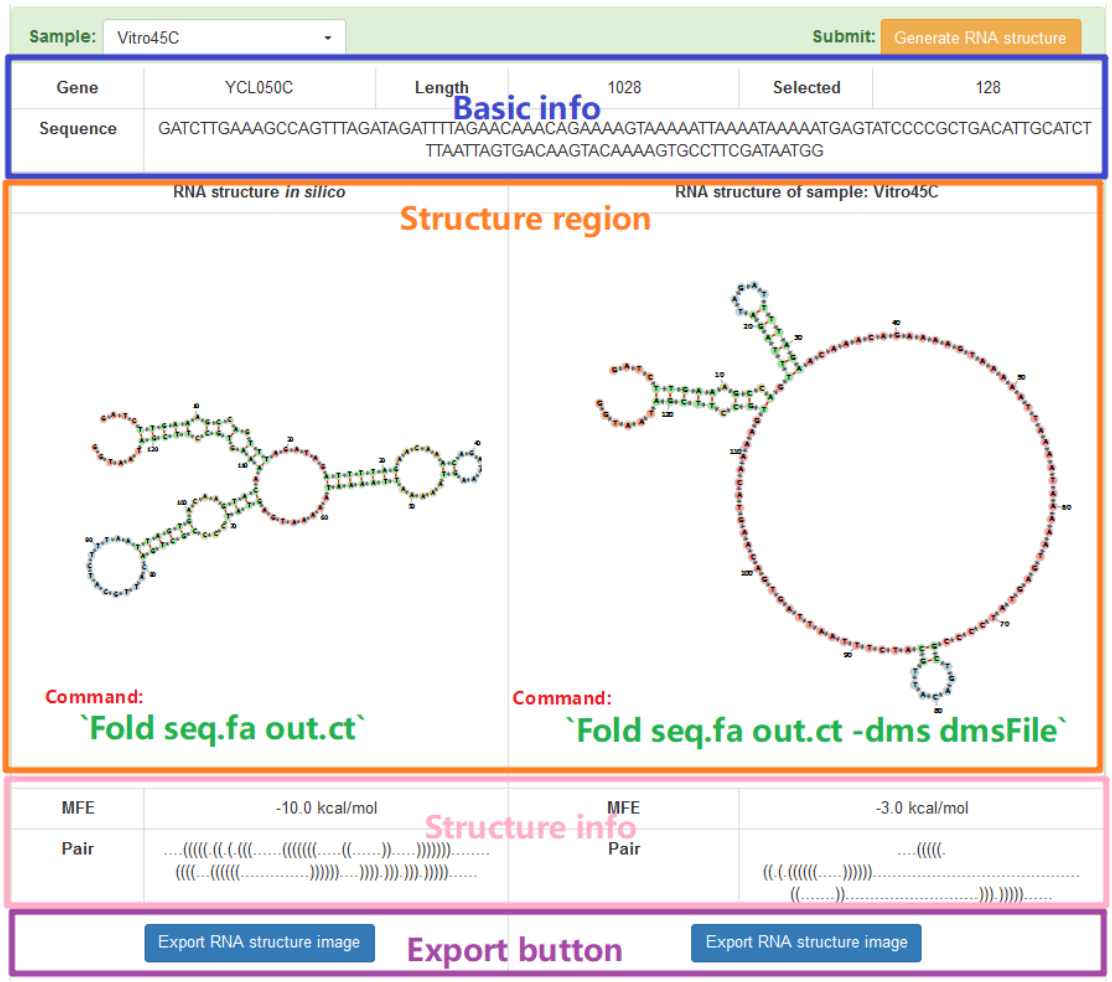
2.4.3.2.2 Predict RNA structure by 'Fold'

We used 'RNAstructure' software of Mathews lab to predict *in silico* RNA structure and experimental structures. RNA structure *in silico* is the RNA structure directly predicted by RNA sequence, while experimental structures are the RNA structure corrected by normalized DMS signal. The code is:

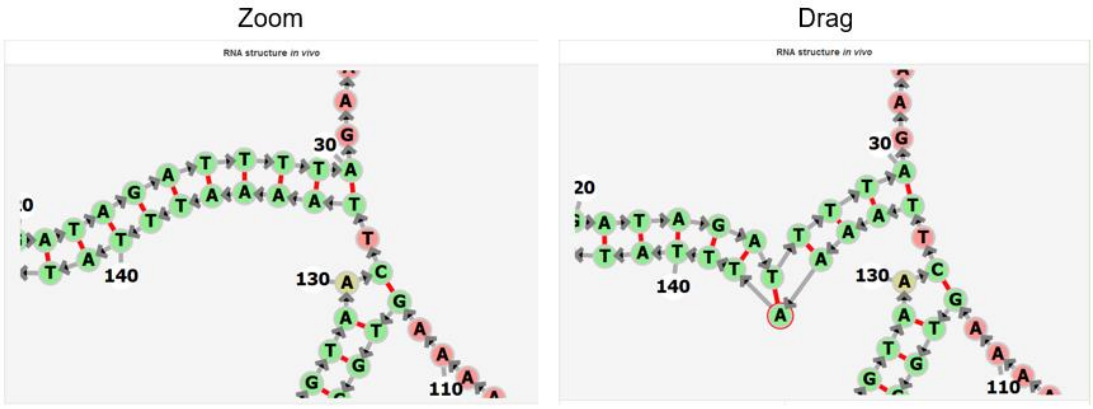
In silico: `Fold seq.fa out.ct`

In experimental: `Fold seq. fa out. ct – dms dmsFile`

Where seq.fa is the RNA sequence, out.ct is the predicted output file, and dmsFile is the file containing the DMS signal. Where, in the DMS signal file, the non-AC site is labeled '-999', which means it is meaningless.

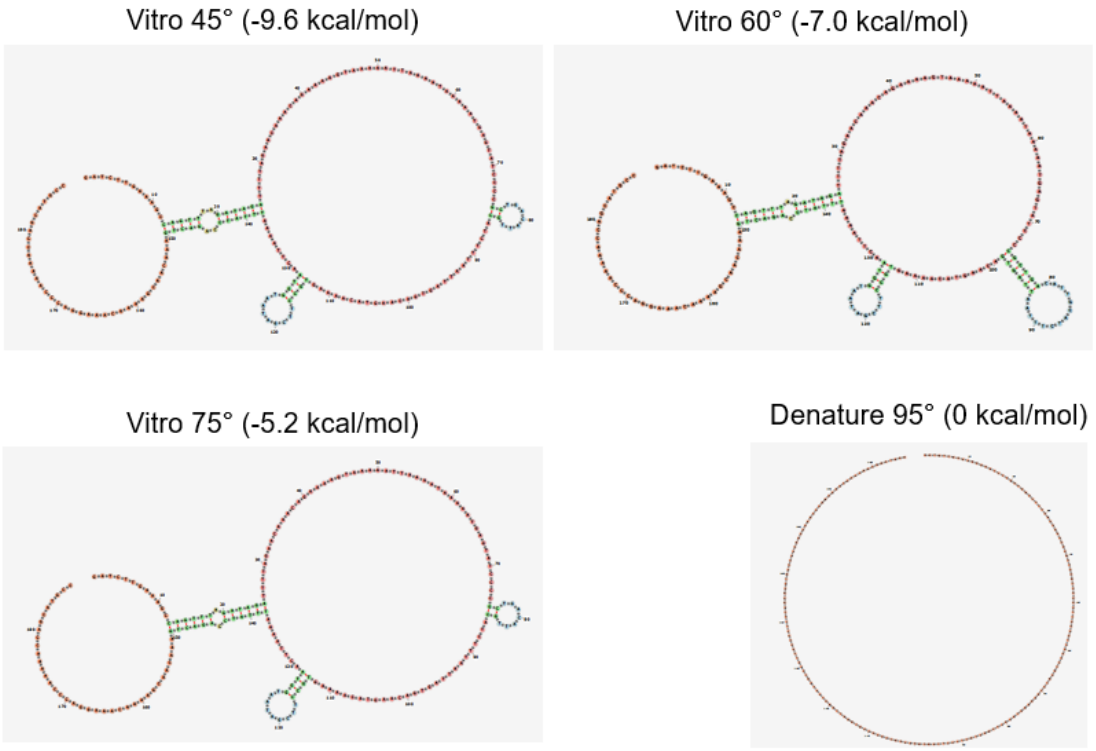


- **Gene: gene id**
- **Length: gene id**
- **Selected:**
- **Sequence: the sequence of the gene**
- **RNA structure *in silico*: visualization of RNA structure *in silico***
- **RNA structure of sample X:visualization of RNA structure in experiment X**
- **MFE: Minimum free energy of the structure**
- **Pair: display the structure with dot and bracket**



RNA structure support zooms and drag

If you want to compare the structure of a gene under multiple experimental conditions, you can change 'Sample' to obtain the RNA structure of the sequence under different experimental conditions.

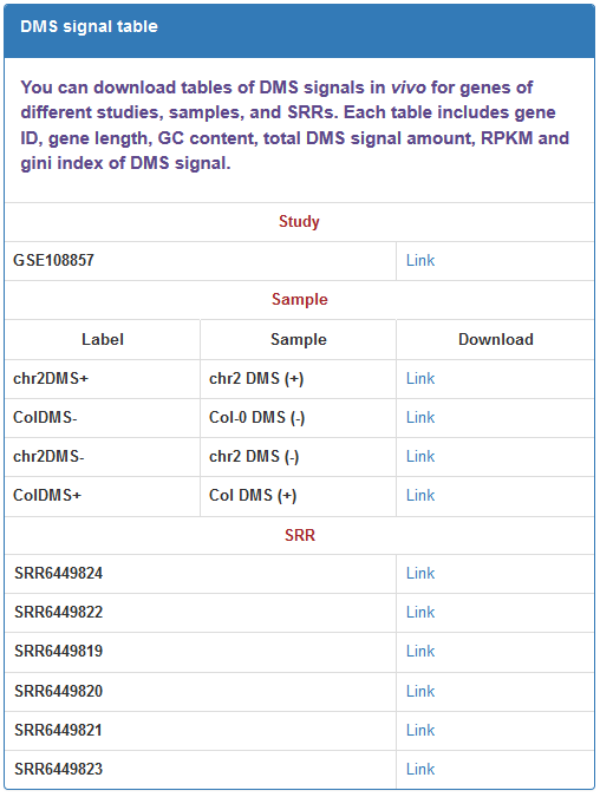


Structure of the same RNA sequence under different experimental conditions.

2.5 Download

The navigation of this page on the left includes species and studies. You can choose one species, and select one study of the species.

2.5.1 DMS signal table



DMS signal table

You can download tables of DMS signals in *vivo* for genes of different studies, samples, and SRRs. Each table includes gene ID, gene length, GC content, total DMS signal amount, RPKM and gini index of DMS signal.

Study	
GSE108857	Link

Sample		
Label	Sample	Download
chr2DMS+	chr2 DMS (+)	Link
ColDMS-	Col-0 DMS (-)	Link
chr2DMS-	chr2 DMS (-)	Link
ColDMS+	Col DMS (+)	Link

SRR	
SRR6449824	Link
SRR6449822	Link
SRR6449819	Link
SRR6449820	Link
SRR6449821	Link
SRR6449823	Link

In this section, we provide download link of tables of DMS signals *in vivo* for genes of different studies, samples, and SRRs. Each table includes gene ID, gene length, GC content, total DMS signal amount, RPKM and Gini coefficient.

- **Total DMS signal amount: sum of DMS raw signal of a transcript**
- **RPKM: Reads Per Kilobase per Million mapped reads(RPKM) of the transcript**
- **Gini coefficient:** After removing the G and T loci, the Gini coefficients of A and C loci of the whole mRNA DMS signals were calculated.

2.5.2 Gini coefficient table

Gini coefficient could be a measure of the strength of mRNA structure (Rouskin et al. 2014). The Gini coefficient of each transcript is calculated by sliding window, with a size of 50 A and C bases (G and T bases are removed) and a step size of 50nt. The table contains 2 columns, gene ID and

Gini coefficient. Different screening thresholds mean that the proportion of 0 sites is higher than the threshold is removed. If the threshold is 100%, no screening will be conducted.

Gini coefficient

Gini coefficient could be a measure of the strength of mRNA structure (Rouskin et al. 2014). The Gini coefficient of each transcript is calculated by sliding window, with a size of 50 A and C bases (G and T bases are removed) and a step size of 50nt. The table contains 2 columns, gene ID and Gini coefficient.Different screening thresholds mean that the proportion of 0 sites is higher than the threshold is removed. If the threshold is 100%, no screening will be conducted.

Label	Sample	0 sites Thresholds					
		50%	60%	70%	80%	90%	100%
chr2DMS+	chr2 DMS (+)	Link	Link	Link	Link	Link	Link
ColDMS-	Col-0 DMS (-)	Link	Link	Link	Link	Link	Link
chr2DMS-	chr2 DMS (-)	Link	Link	Link	Link	Link	Link
ColDMS+	Col DMS (+)	Link	Link	Link	Link	Link	Link

Note that the Gini coefficient here is different from the Gini coefficient in the previous table.The Gini coefficient in this table is carried out by sliding form (after removing G and T, the length of form is 50, and the sequence length is about 100).