RSVdb: A comprehensive database of *in vivo* mRNA structure

Haopeng Yu[1,2,†], Yi Zhang[1,2,†], Qing Sun[1,2], Huijie Gao[3] and Shiheng Tao[1,2,*]

[1] College of Life Sciences and State Key Laboratory of Crop Stress Biology in Arid Areas, Northwest A&F University, Yangling, Shaanxi, 712100, China
[2] Bioinformatics Center, Northwest A&F University, Yangling, Shaanxi, 712100, China
[3] College of Animal Science and Technology, Northwest A&F University, Yangling, Shaanxi, 712100, China

* To whom correspondence should be addressed. Tel: +86 029-87082976; Email: shihengt@nwsuaf.edu.cn

† These authors contributed equally to the paper as first authors.

ABSTRACT

RNA performs a crucial regulatory role in cells by folding into complex RNA structures. To date, chemical compounds have been developed to effectively probe mRNA structure *in vivo* at the transcriptome level. Here, we introduce a database, RSVdb (https://taolab.nwafu.edu.cn/rsvdb/), to browse and visualize mRNA structure data *in vivo*. RSVdb, including 622,429 RNAs from 178 samples in 10 studies of 8 species, provides four main functions: information retrieval, research overview, structure prediction, and resource download. Users can search for species, studies or genes of interest through fuzzy search and search suggestions; browse sequencing data quality control information and statistical charts of mRNA structure; predict and visualize RNA structure *in silico* and in experimental treatments; and download RNA structure data for the available species and studies. Overall, RSVdb provides an easy reference for RNA structure at the transcriptome level and will support future research on the functions of RNA structure in cells.

INTRODUCTION

Intracellular mRNA is not only the carrier of genetic information but also the regulator of translation initiation, ribosome translation rate, subcellular mRNA localization, and cotranslational protein folding by forming mRNA structures *in vivo* (1-6). Several compounds have been developed to detect RNA structures *in vivo*. For example, dimethyl sulfate (DMS) can identify non-Watson-Crick conformations of adenine and cytosine, and information about mRNA structure *in vivo* can then be calculated through the RNA structure algorithm (7-9). DMS probing methods use next-generation sequencing technology to monitor the structure of mRNA, and the first global detection of *Arabidopsis thaliana*, *Saccharomyces cerevisiae,* and *Homo sapiens* mRNA structures has been performed *in vivo* (10,11). This method can effectively detect the structure of mRNA *in vivo* at the transcriptome level. In recent years, an increasing number of studies have used this method or improved techniques, such as structure-seq, CIRS-seq, and DMS-MaPseq, to detect mRNA structures (12-18). Currently, there are several databases collecting *in vivo* mRNA structure information: RMDB (https://rmdb.stanford.edu/browse/), FoldAtlas (http://www.foldatlas.com/), and Structure Surfer (http://tesla.pcbi.upenn.edu/structuresurfer/)(19-21). RMDB contains the chemical mapping data of

specific RNA structures in different experimental treatments rather than *in vivo* mRNA structures at the transcriptome-wide level. FoldAtlas contains the *in vivo* mRNA structure of *Arabidopsis thaliana* labeled by DMS, and Structure Surfer includes only the *in vivo* RNA structure data of *Homo sapiens* and *Mus musculus*.

Here, we collected all the studies of *in vivo* mRNA structure labeled by DMS, and we present a comprehensive, visual and user-friendly database, RSVdb (https://taolab.nwsuaf.edu.cn/rsvdb/). RSVdb. It includes 622,429 RNAs from 178 samples in 10 studies of 8 species, statistical presentation of mapping data, and even prediction and visualization of mRNA structure *in silico* and in experimental treatments (Figure 1B). The database provides interactive charts showing statistical analyses at the transcriptome, sample, and gene levels, including sequencing data quality control, DMS signal distribution, Gini coefficient distribution of the DMS signal, and more. Furthermore, it allows users to predict and view the structure of the mRNA of interest based on the available DMS signals. In addition, DMS signal data and Gini coefficient data under different thresholds can be downloaded. All charts in the database are interactive and can be exported online as figures or in text format, and all search boxes in the database support fuzzy search and search suggestions (Figure 1). In summary, RSVdb makes it easy to browse, search, and download mRNA structure data.

MATERIAL AND METHODS

Data sources

The DMS sequencing data were collected from the Gene Expression Omnibus (GEO) and Short Read Archive (SRA) databases. The current version contains 178 samples from 10 studies of 8 species. The sequencing data of these samples were mapped to the corresponding reference genomes and transcriptomes and then normalized through data processing to screen 622,429 RNAs with valid data (Figure 2).

Data processing

Sequencing data mapping. All SRA files were obtained from the NCBI SRA database and converted to Fastq format by using 'pfastq-dump' (https://github.com/inutano/pfastq-dump), which requires the installation of 'sratoolkit' v2.9.6 (https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software). Quality control checks on the raw sequence data were performed by FastQC v0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc). For Fastq files, the read adapters were removed by Cutadapt v1.18 (https://cutadapt.readthedocs.io/en/stable/index.html) with the parameters '-j 8 -n [N] -b'. All reads were filtered by FASTX-Toolkit v0.0.14 (http://hannonlab.cshl.edu/fastx_toolkit/download.html) with a quality score higher than 30 and a read length greater than 21. The reference genomes of the eight species and corresponding annotated files are shown in Table S1. The transcriptome sequences of the eight species were downloaded from Ensembl, and only the longest transcript was retained for each gene. Hisat2 v2.1.0 (https://ccb.jhu.edu/software/hisat2/index.shtml) was used for sequence alignment, and two mismatches were allowed. We mapped the Fastq files to both the reference genome and reference

transcriptome. It is worth mentioning that reads may map to the untranscribed regions in the genome, thus reducing the read abundance of the transcriptome. Therefore, based on the comparison with genome mapping statistics, we recommend mapping to the transcriptome directly.

Normalization. We performed 2%/8% normalization on the raw DMS signal (22). In this step, the top 2% of DMS signal values are treated as outliers, and each DMS signal value (including the top 2%) is divided by the average of the remaining top 8% of DMS signal values to obtain normalized structural reactivity. Transcripts with no DMS signal in the 2% to 8% range were removed.

RNA structure prediction. RNA structures *in silico* and in experimental treatments were predicted by 'Fold' of 'RNAstructure' software package (7).

Data analysis

We used Perl scripts to generate three types of statistics for the mapped reads: (i) the proportion of mapped and unmapped reads in each study, (ii) the numbers of mapped and unmapped reads in each study and (iii) the mapping ratios of the four bases A, C, G and T. Then, we calculated two additional types of statistics for the mapped reads: (i) the distribution of DMS RPKM values and (ii) the distribution of the Gini coefficients of the 5 'UTR, CDS, and 3' UTR of each sample.

RPKM was used as an index to measure the abundance of DMS signal for each transcript, as defined below.

$$\mathrm{RPKM} = \frac{r_t \times 10^9}{R \times fl_t}$$

where r_t is the raw DMS count in a transcript, R is the total mapped reads in all transcripts, and fl_t is the feature length of a transcript.

The Gini coefficient was used as an index to measure the strength of the mRNA structure for each transcript, which was calculated using a sliding window with a size of 50 A and C bases (G and T bases were removed)(11,23). The Gini coefficient is defined as follows.

$$G_j = 1 - \frac{1}{n}(2\sum_{i=1}^{n-1} W_i + 1)$$

$$\mathrm{G} = \frac{1}{m}\sum_{j=1}^{m} G_j$$

where W_i is the cumulative DMS count from the first base to the *i*th base as a percentage of the total DMS count in the window j. n is the total number of A and C bases in a window, which was 50. G_j is the Gini coefficient of the *j*th window, and m is the total window count of the transcript. If the DMS count in a window is zero, the Gini coefficient of the window is zero.

Front-end interface and back-end structure

The implementation of the RSVdb database can be divided into two parts: the front end (client side) and back end (server side). The back end implementation uses the python-flask web framework, which is responsible for server-side website operation logic, necessary data processing, and SQLAlchemy, which is responsible for data storage and query. The front-end interface is written in HTML 5, CSS, and JavaScript and uses AJAX to asynchronously interact with the data on the server side. The front-end and back-end code of RSVdb are freely available via the GitHub repository (https://github.com/atlasbioinfo/RSVdb/).

RESULTS

Database usage and access

The RSVdb website contains four functions: information retrieval, research overview, structure prediction, and resource download.

Information retrieval. Our website provides a powerful retrieval function. On the ‘Home’ page, the top search box can search for species, GEO or SRA numbers. In the ‘Quickstart’ section below, users can directly click a picture of a species or functional module to enter the corresponding page. Meanwhile, the viewing page search box also offers searches for the transcript IDs and gene names of 622,429 RNAs.

Research overview. Our website provides the necessary experimental information and statistics for each study. On the 'Browse' page, users can click the species and research bar on the left to obtain information on the corresponding research. The 'Study info' section lists the necessary information on the study, in which the QC section gives an HTML report on the data quality control. The section 'Transcriptome and genome mapping statistics' gives significant differences in mapped read counts between the transcriptome and genome, as well as the proportions of reads for the four bases A, C, G, and T. The ‘Sample’ section lists the samples and SRA numbers for each study. The section ‘Sample mapping statistics’ shows the differences among the samples. The section ‘Histogram of DMS RPKM’ plotted the frequency histogram of RPKM of DMS signal in each transcript under different experimental conditions. The section ‘Boxplot of Gini coefficient’ compares the Gini coefficient of 5 'UTR, CDS, and 3' UTR in each transcript among different samples. 'Sample info' section lists the experimental processing information for the different samples. The section 'Histogram of Gini coefficient' plots the frequency histogram of the Gini coefficient for the 5 'UTR, CDS, and 3' UTR of each transcript. Finally, the ‘RPKM top 100 transcripts’ section lays out the essential information of the first 100 genes ranked by RPKM as calculated by DMS signal.

Structure prediction. Our website provides online structure prediction and interactive display. On the 'Viewer' page, enter any transcript ID or gene name from the selected species in the search box to obtain the statistical information of the sequence and an interactive bar plot of the normalized DMS signal. Subsequences of arbitrary length can be selected to generate predicted structures *in silico* and

DMS corrected structures under different experimental conditions, and all generated structures can be presented interactively.

Resource download. All the generated images and RNA structure data of each transcript in different experimental conditions are available for download from our website.

RSVdb also provides detailed illustrations of database usage and functional demonstrations. Complete user manuals for RSVdb are also available online and for download.

DISCUSSION

RSVdb is a comprehensive and user-friendly *in vivo* mRNA structure database with multiple functions. It not only supports browsing data related to mRNA structure *in vivo*, including sequencing data quality control information, statistical display of DMS signal data, and statistical display of Gini coefficient, but also predicting and visualizing mRNA structure *in silico* and in experimental treatments. The database also allows users to download the available *in vivo* mRNA structure data, so interactive online manipulation and export of charts are supported, and detailed user manuals are provided. RSVdb enables users to easily access multiple platforms through the browser, with additional optimizations for mobile platform access and charting operations. Additionally, we provide the open-source code used to build the front and back ends of the RSVdb in the GitHub repository, which supports downloading and building the database locally to achieve its main functions offline.

Although RSVdb has certain advantages, there are still some limitations. Currently, the database contains only *in vivo* mRNA structure data from labeling with DMS reagent, but excellent and continuously improving new methods, such as SHAPE and CMCT, are becoming available to label *in vivo* mRNA structure (3,24,25). In the future, we will continue updating the database by adding *in vivo* mRNA structure data labeled with SHAPE, CMCT, and other reliable reagents. Moreover, we will work to combine these *in vivo* structural data with other data, such as ribosome profiles, to show the interaction between mRNA structure and ribosomes during translation.

Acknowledgment

The authors would like to thank the Network & Education Technology Center of NWAFU for its server hardware and network services support. We are grateful to Shuo Gao, Xuanyan Li, and Jingjing Son for their advice on this project in terms of server operation and webpage technology.

FUNDING

This work was supported by the National Natural Science Foundation of China (Grant 31771474).

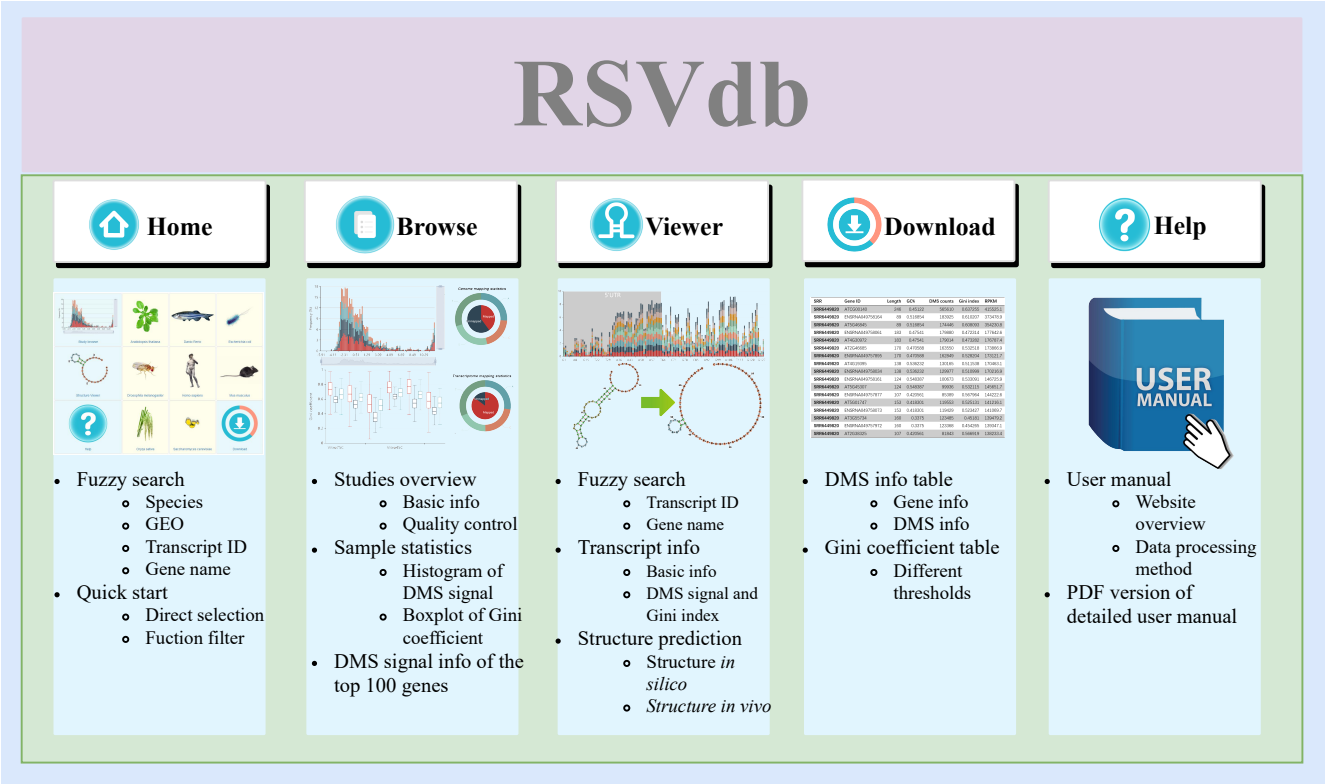
REFERENCES

1. Lecuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., Hughes, T.R., Tomancak, P. and Krause, H.M. (2007) Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell*, **131**, 174-187.
2. Martin, K.C. and Ephrussi, A. (2009) mRNA Localization: Gene Expression in the Spatial Dimension. *Cell*, **136**, 719-730.
3. Mustoe, A.M., Busan, S., Rice, G.M., Hajdin, C.E., Peterson, B.K., Ruda, V.M., Kubica, N., Nutiu, R., Baryza, J.L. and Weeks, K.M. (2018) Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing. *Cell*, **173**, 181-195 e118.
4. Espah Borujeni, A., Cetnar, D., Farasat, I., Smith, A., Lundgren, N. and Salis, H.M. (2017) Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Res*, **45**, 5437-5448.
5. Faure, G., Ogurtsov, A.Y., Shabalina, S.A. and Koonin, E.V. (2016) Role of mRNA structure in the control of protein folding. *Nucleic Acids Res*, **44**, 10898-10911.
6. Yu, H.P., Meng, W.J., Mao, Y.H., Zhang, Y., Sun, Q. and Tao, S.H. (2019) Deciphering the rules of mRNA structure differentiation in Saccharomyces cerevisiae in vivo and in vitro with deep neural networks. *Rna Biology*.
7. Mathews, D.H. (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *Rna*, **10**, 1178-1190.
8. Wells, S.E., Hughes, J.M.X., Igel, A.H. and Ares, M. (2000) Use of dimethyl sulfate to probe RNA structure in vivo. *Method Enzymol*, **318**, 479-493.
9. Cordero, P., Kladwang, W., VanLang, C.C. and Das, R. (2012) Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry*, **51**, 7037-7039.
10. Ding, Y., Tang, Y., Kwok, C.K., Zhang, Y., Bevilacqua, P.C. and Assmann, S.M. (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, **505**, 696-700.
11. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. and Weissman, J.S. (2014) Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, **505**, 701-705.
12. Burkhardt, D.H., Rouskin, S., Zhang, Y., Li, G.W., Weissman, J.S. and Gross, C.A. (2017) Operon mRNAs are organized into ORF-centric structures that predict translation efficiency. *Elife*, **6**.
13. Ritchey, L.E., Su, Z., Tang, Y., Tack, D.C., Assmann, S.M. and Bevilacqua, P.C. (2017) Structure-seq2: sensitive and accurate genome-wide profiling of RNA structure in vivo. *Nucleic Acids Res*, **45**, e135.
14. Incarnato, D., Neri, F., Anselmi, F. and Oliviero, S. (2014) Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian transcriptome. *Genome Biol*, **15**.
15. Zubradt, M., Gupta, P., Persad, S., Lambowitz, A.M., Weissman, J.S. and Rouskin, S. (2017) DMS-MaPseq for genome-wide or targeted RNA structure probing in vivo. *Nat Methods*, **14**, 75-82.
16. Wang, Z., Ma, Z., Castillo-Gonzalez, C., Sun, D., Li, Y., Yu, B., Zhao, B., Li, P. and Zhang, X. (2018) SWI2/SNF2 ATPase CHR2 remodels pri-miRNAs via Serrate to impede miRNA production. *Nature*, **557**, 516-521.
17. Guenther, U.P., Weinberg, D.E., Zubradt, M.M., Tedeschi, F.A., Stawicki, B.N., Zagore, L.L., Brar, G.A., Licatalosi, D.D., Bartel, D.P., Weissman, J.S. *et al.* (2018) The helicase Ded1p controls use of near-cognate translation initiation codons in 5' UTRs. *Nature*, **559**, 130-134.
18. Beaudoin, J.D., Novoa, E.M., Vejnar, C.E., Yartseva, V., Takacs, C.M., Kellis, M. and Giraldez, A.J. (2018) Analyses of mRNA structure dynamics identify embryonic gene regulatory programs. *Nat Struct Mol Biol*, **25**, 677-686.
19. Yesselman, J.D., Tian, S.Q., Liu, X., Shi, L., Li, J.B. and Das, R. (2018) Updates to the RNA mapping database (RMDB), version 2. *Nucleic Acids Res*, **46**, D375-D379.

20. Norris, M., Kwok, C.K., Cheema, J., Hartley, M., Morris, R.J., Aviran, S. and Ding, Y.L. (2017) FoldAtlas: a repository for genome-wide RNA structure probing data. *Bioinformatics*, **33**, 306-308.
21. Berkowitz, N.D., Silverman, I.M., Childress, D.M., Kazan, H., Wang, L.S. and Gregory, B.D. (2016) A comprehensive database of high-throughput sequencing-based RNA secondary structure probing data (Structure Surfer). *Bmc Bioinformatics*, **17**.
22. Ding, Y.L., Kwok, C.K., Tang, Y., Bevilacqua, P.C. and Assmann, S.M. (2015) Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. *Nat Protoc*, **10**, 1050-1066.
23. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, **5**, 621-628.
24. Low, J.T. and Weeks, K.M. (2010) SHAPE-directed RNA secondary structure prediction. *Methods*, **52**, 150-158.
25. Frezza, E., Courban, A., Allouche, D., Sargueil, B. and Pasquali, S. (2019) The interplay between molecular flexibility and RNA chemical probing reactivities analyzed at the nucleotide level via an extensive molecular dynamics study. *Methods (San Diego, Calif.)*.

TABLE AND FIGURES LEGENDS

A



B

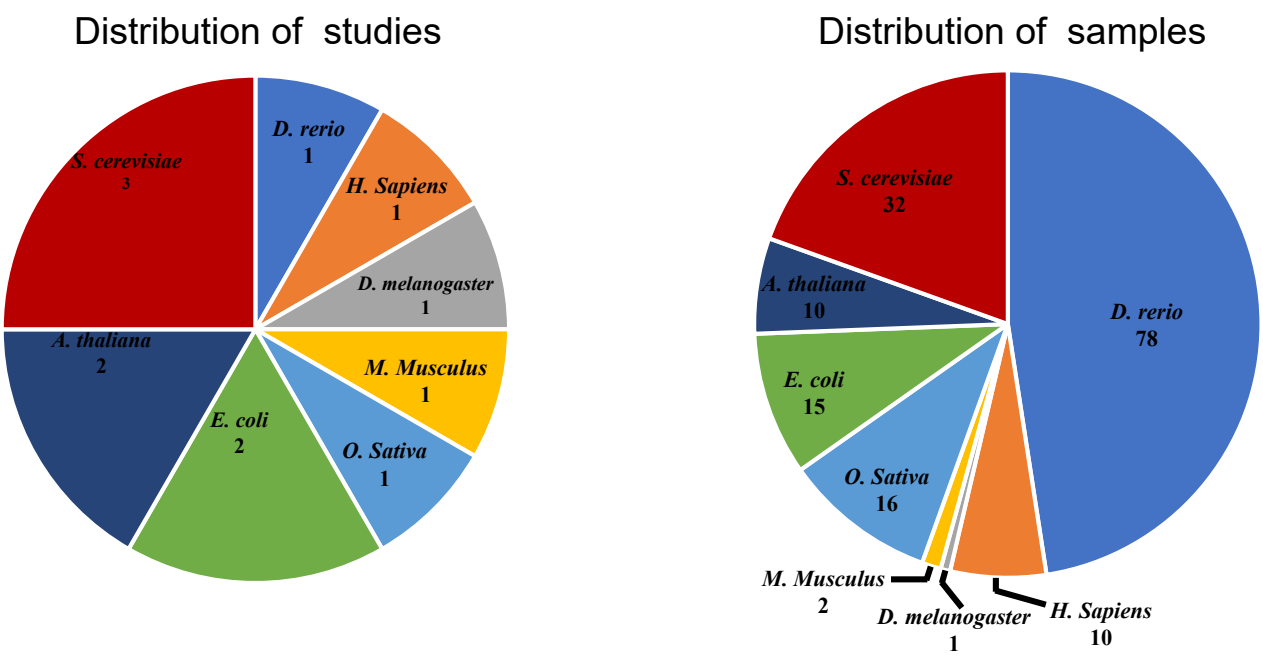


Figure 1. Schematic overview and main contents of RSVdb. (A) Website frame and main functions. (B) Distribution of studies and samples among species.

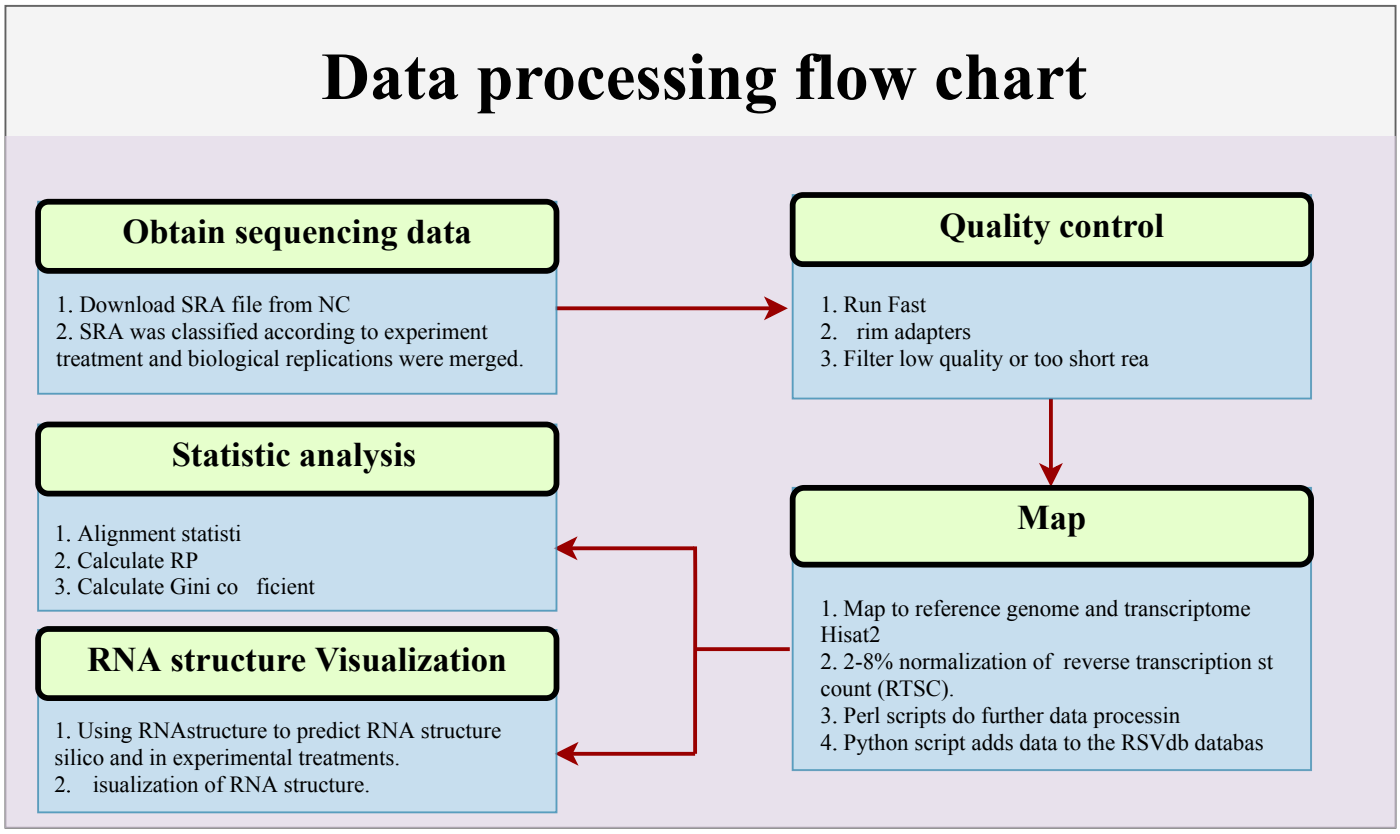


Figure 2. Data processing flow of RSVdb.