

ac4C Info: A database to parse N4-acetylocytidine(ac4C) epitranscriptome in multiple species

Kang Zhang

zk7588@163.com

Da Sun

Hebei Agricultural University

Zimu Li

Jihong Xing

Jingao Dong

Hebei Agricultural University

Article

Keywords: RNA modifications, N4-acetylcytidine (ac4C), Database, Epitranscriptome, Deep learning, SNP, Multiple species

Posted Date: October 23rd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10908712/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Abstract

In RNA modification research, the functions, molecular features, and mechanisms behind RNA acetylation are mostly unknown. In order to better understand the transcriptome-wide distribution patterns of ac4C modifications among various species as well as the related biological implications, we have created the ac4C Info platform, which includes four main modules. The ac4CDB database integrates two types of data: (I) single-base resolution mapping of ac4C RNA acetylation sites identified by ac4C-seq technology, and (II) transcriptome-wide profiling of ac4C distribution regions derived from acetylated RNA immunoprecipitation sequencing (acRIP-seq). ac4CDiseaseDB database gathers disease-related ac4C SNPs – those which result in the acquisition or removal of ac4C sites – from two sources: germline variants obtained from dbSNP, the 1000 Genomes Project, and Ensemble; and somatic variants found in 33 different human cancers collected from The Cancer Genome Atlas (TCGA). Additionally, there are two analytical modules included in this platform: one is an ac4C predictor that can make real-time, interactive predictions about ac4C modification sites; the other is an ac4C SNP IFer used to evaluate how genetic variations affect ac4C modifications. We anticipate that ac4C Info will serve as a comprehensive and valuable resource platform, enabling data sharing, centralized management, functional annotation, and customized analysis of ac4C-related information. It is freely accessible at: <https://hebaubioinformatics.cn:447/>.

1. Introduction

With the birth of epitranscriptomics, RNA modifications have been recognized as the main player in the regulatory network of gene expression, which has become an important advance to understand the fine-tuned mechanisms of life process in post genomic era. RNA modifications as a major epigenetic mechanism in post-transcriptional gene regulation, are heavily engaged in basic cell functions like proliferation, differentiation and apoptosis through affecting RNA splicing, transportation, longevity and translation rate. Dysregulation of these modifications is closely associated with the onset and progression of various diseases, such as cancers, neurodegenerative disorders, and metabolic syndromes.

Of all the more than 170 different kinds of RNA modifications that have been discovered so far, N4-acetylcytidine (ac4C) is high evolutionarily conserved acetylation modification. It can be found in many different types of RNA like mRNA, tRNA, rRNA in plants and animals and bacteria. Owing to its important role in regulating gene expression, ac4C has emerged as a focal point of research in the field of epitranscriptomics. In recent years, with the continuous innovation of high-throughput sequencing technologies such as next-generation sequencing (NGS) and Oxford Nanopore Technology (ONT), the detection technology of ac4C modification has developed rapidly. Targeted sequencing methods such as acRIP-seq, as well as ac4C-seq technology based on chemical labeling (which can be combined with NGS or directly used on RNA sequencing platforms) and ONT direct RNA sequencing technology have been established successively, providing strong technical support for systematically analyzing the distribution characteristics and biological functions of ac4C modification in the entire transcriptome.

Relying on these technologies, researchers have identified a large number of ac4C modification sites in different species, tissue types, and pathological states, gradually revealing its extensive functions in physiological and pathological processes.

In the field of embryonic development, NAT10-mediated ac4C modification can maintain the self-renewal ability of human embryonic stem cells by stabilizing the mRNA of core pluripotency transcription factor OCT4, thereby participating in the regulation of early mammalian embryonic development^[1]; in disease-related research, ac4C modification can mediate spinal cord central sensitization by enhancing the translation efficiency of Vegfa mRNA, and participate in the occurrence and development of neuropathic pain^[2]; at the same time, it can also promote tumor progression by regulating the expression of oncogenes, and thus is regarded as a potential disease diagnostic marker and therapeutic target^[3]; in the field of plant biology, ac4C modification is widely distributed in chloroplast transcripts and photosynthesis-related genes, and can affect the photosynthesis process of plants by regulating the translation efficiency of genes related to the light reaction transmission chain^[4].

The above-mentioned research results clearly show that ac4C modification is very important for many different cellular processes and at the same time it has caused a lot of ac4C related omics data to accumulate quickly. Although we have made great progress in the field of ac4C modification, it is still difficult to integrate and deeply explore such large amount of data. Firstly, the existing ac4C modification data is scattered among many research papers, public databases, and individual research teams without a unified data integration platform and standard annotation system. Heterogeneity in detection methods and data analysis pipelines between studies causes different data formats with low comparability which greatly hinders thorough cross-study, cross-species, and cross-pathological condition analysis.

Secondly, the current data coverage is very limited, most of the data are from a small number of model organisms like human and mouse, there is no enough data for non-model species such as plants and prokaryotes. And also, ac4C modification data obtained by direct RNA sequencing with Oxford Nanopore Technologies (ONT) has not been systematically incorporated, this makes it hard to build an all-around ac4C modification picture. Most importantly is that we do not know what the functional mechanisms of ac4C modifications are. Scientific questions such as its connection to disease-related genetic variants, the topology of regulatory networks, and whether there will be any cooperative or opposing relationships with other RNA modifications, all need more research and support through multi-data integration and advanced analysis methods. As it stands now, the majority of epitranscriptome databases tend to concentrate on better known RNA modifications such as m6A^[5, 6], m5C^[7], m7G^[8], whereas dedicated resources for ac4C are extremely rare and fail to satisfy the increasing need for specialized data infrastructure in this new area.

In such an environment, the creation of a thorough and organized database that is centered around ac4C RNA modifications has both significant scientific value and real-world application possibilities. We would like to create a one-stop ac4C RNA modification database by systematically integrating ac4C modification data obtained from different experimental platforms (acRIP-seq, ac4C-seq) across many

species including animals, plants, and fungi (it will be updated soon) and many diseases. Database will give standard annotation info like sequence features of mod site, Functional annotation of target genes and their association with diseases. And also help with quick and clear searching and showing of ac4C change spots and have all kinds of mixed data analyzing tools to do more detailed looking at how ac4C changes work and if they can be used in medicine. The database will be created which will fill the gap of dedicated ac4C data resource and make the data sharing and usage easy and efficient, so that we can give high quality data support to the global ac4C research community and promote the research and application in physiological regulation, disease mechanism and crop genetic improvement.

2. Material and methods

2.1. Collection and Analysis of ac4C RNA modifications datasets

The ac4C sites collected in the ac4C Info are derived from two mainstream detection techniques used to analyze N4-acetylcytosine (ac4C) RNA modifications: ac4C-seq and acRIP-seq. The ac4C sites were obtained from the analysis of 168 sequencing samples using these two different ac4C analysis techniques. Sequencing samples were collected from 35 independent studies in the NCBI SRA database, CNCB database^[9], and ENA database (<https://www.ebi.ac.uk/ena/browser/home>), including 332 FASTQ files. Among them, there were 4 groups of ac4C-seq data and 328 groups of acRIP-seq data. The detailed dataset of the ac4C modification sites identified in this study can be found in (Supplementary Table 1).

2.2. Analysis of ac4C RNA modifications datasets

NGS technology (base resolution): The ac4C sites categorized within this group were identified at single-base resolution from NGS-based ac4C-seq studies^[1]. These sites were extracted from the corresponding supplementary materials of the respective ac4C-seq publications. NGS technology (including ac4C-modified regions): For the processing of ac4C-RIP-seq data, fastp software (v0.23.4)^[10] is first employed to remove sequencing adapters and filter out low-quality reads, thereby generating high-quality clean FASTQ reads. Subsequently, clean reads are aligned to the corresponding reference genome using HISAT2 (v2.2.1)^[11], with the same version of genome annotation adopted as the reference gene model. The resulting SAM-format alignment files are then sorted, converted to BAM format, and indexed using samtools (v1.6)^[12]. Finally, R package exomePeak2 software (v1.14.3) is applied to analyze the BAM-formatted reads from each paired IP + input (or IgG + input) sample to identify significant acetylation peaks in the IP or IgG samples. ExomePeak2 is a specialized tool designed for RNA modification peak calling, capable of accurately detecting modification sites from splicing-aware RNA-seq data by incorporating known exon-intron structures^[13]. The analysis is performed using default parameters and the provided reference gene model. For datasets that include an IgG control group, non-specific binding signals are accounted for by removing any IP-derived peaks that overlap more than 10% with IgG peak

coordinates, achieved through the subtraction function in bedtools (v2.31.1)^[14]. All identified sites are annotated using custom Python and R scripts. Annotation details include: chromosome (Chr), species, site and peak position information, the associated gene, the core structural region of the RNA corresponding to the site, BioType (biological type annotation of the gene or transcript), tissue or cell line origin of the sample, and the nucleotide sequence encompassing the peak region. Intersect the potential involvement data of post-transcriptional regulation collected using POSTAR2 (RBP binding regions)^[15] and starBase (miRNA–RNA targets)^[16] with the identified site information, and record the overlapping results.

2.3. Code integration and training optimization for ac4C modification site prediction under a multi-model framework

This project is devoted to the systematic incorporation and performance improvement of current computational models for forecasting N4-acetylcytosine(ac4C) modification sites in RNA sequences. To increase prediction accuracy and model robustness, we have systematically gathered and incorporated nine publicly available model frameworks from separate research endeavors: LSA-ac4C^[17], TransAC4C^[18], ERNIE-ac4C^[19], Voting-ac4C^[20], DLC-ac4C^[21], DeepAc4C^[22], NBCR-ac4C^[23], STM-ac4C^[24], and TransCNN^[25]. All of these above models have been created especially for finding out the N4-acetylcytosine(ac4C) modification site in RNA, they all use different main structures and ways of doing things. LSA-ac4C creates a hybrid neural network using a two-layer LSTM network and a self-attention mechanism. TransAC4C is based on the Transformer encoder and uses bidirectional LSTM together with a one-dimensional convolutional network (Conv1D) to make it work for many species and predict single-base resolution. ERNIE-ac4C uses the ERNIE-RNA pre-trained language model to obtain the sequence level semantic features and attention map, and then realizes the multi-level feature fusion through a 2D CNN. Voting-ac4C uses the RNAErnie pre-trained model and adds six other common feature encodings (One-hot, ENAC) to create a high-dimensional feature space. The final prediction is done by a soft voting ensemble after dimensionality reduction using a deep learning network. DLC-ac4C has a DenseNet and BiLSTM architecture, it includes a channel attention mechanism and uses an isomorphic ensemble method to improve the model's generalizability. DeepAc4C is based on a 1D-CNN framework and uses hybrid feature input which consists of rational attribute pattern and nucleic acid distributed representation. NBCR-ac4C uses the Nucleotide Transformer and DNABERT2 models to create context-aware nucleotide embeddings and then combines CNNs and ResNet18 for both shallow and deep semantic features. TransCNN combines the Transformer architecture and a text convolutional network (TextCNN) using an adaptive feature extraction method and multiple scale filters to obtain both local and global context information from RNA sequences. STM-ac4C uses a special kind of deep learning structure so it can find ac4C sites well (Supplementary Table 2).

All of the original code repositories, technical documentation, training protocols, and related hyperparameter configurations for all of the models were obtained from publicly available code sharing

platforms like GitHub. We have used only the benchmark dataset, training set and independent test set which was clearly mentioned or cited in their own original paper during our whole research process. All models were retrained under the same hardware and software conditions so that experiments could be reproduced and compared across different models fairly. The training procedure followed very closely the main parameter settings reported by each original study, including the learning rate, batch size, optimizer type, number of training epochs, and regularization techniques, thus retaining as much as possible the intended design and performance characteristics of the model. After the retraining and validation evaluation was completed, all nine models were deployed onto a central server infrastructure that had been set up with GPU acceleration so as to allow for fast parallel computation. This deployment architecture allows external access through standard network service interfaces which lets researchers send in RNA sequences for analysis at the same time receiving ac4C site predictions from many different models. The proposed multi-model integration framework is able to help with consensus based prediction analysis as well as allow users to fully assess the prediction confidence and reliability by looking at the output from different algorithms.

This project brings together all those separate and isolated model developments into one single, cohesive, and easily accessible computational environment to greatly improve how practical, reproducible, and applicable the ac4C modification site prediction tool is. This platform can make sure that we have a good and fast way to know what these possible ac4C places do, which helps us learn more about living things and their genes in the future, and also makes it easier for us to understand how tiny changes in RNA can change how our bodies work.

2.4. Evaluation of the Impact of Genetic Variations on the ac4C Acetylation State in the Epitranscriptome

We looked at how two kinds of genetic changes, called germline and somatic mutations, affect something called ac4C acetylation in the epitranscriptome. Germline variation data was taken from public databases such as dbSNP(v1.55)^[26], the 1000 Genomes (Phase 3), and Ensemble (Release 115)^[27]. We got somatic mutation information for 33 different sorts of human cancers from The Cancer Genome Atlas (TCGA)^[28]. In order to figure out what happens when there's a single nucleotide polymorphism (SNP) in relation to ac4C changes, we used LSA-ac4C, which was a deep neural network model made in another research project, as our main algorithm. This model can learn sequence-based features by itself from mRNA transcripts and classify possible ac4C sites through a hybrid structure combining an embedding module, a two-layer LSTM network, and a self-attention mechanism. One big plus about LSA-ac4C is that you don't have to do any manual work to create your own features and at the same time, it can still get the job done well when it comes to looking at both small and large parts of the sequence around something. When compared with current state-of-the-art models, it shows better accuracy and computation speed, allowing for high throughput ac4C site prediction. Use this model together with some custom python scripts to describe genetic variants according to how they might change ac4C modification maps – specifically whether they cause more or fewer ac4C sites. We set up three levels of certainty for sorting out these variant impacts: (I) High: Genetic variations result in the

gain or loss of modified nucleotides, with an association score for ac4C modification change exceeding 0.8; (II) Medium: The association score for ac4C modification gain or loss ranges between 0.6 and 0.8; (III) Low: The association score falls between 0.5 and 0.6.

The prediction sites with association levels less than 0.5 were excluded from further analysis because they were not reliable. We found out some variants which we called ac4C loss event or ac4C gain event. These variants were annotated by the previously mentioned method and was added with more details like nucleotide change before and after the variant, if it is a variant, its association level, confidence level and so on. And also we got lots of disease associated variants(TagSNP) from the ClinVar database^[29]. These disease-related TagSNPs were then mapped to all ac4C-associated variants to investigate the potential role of ac4C acetylation in disease pathogenesis.

2.5. Implementation of the Database and Web Interface

MySQL database management system is used to store and manage all data sets in ac4C Info. Database querying functions and the user interface were created with PHP and JavaScript. Layout and rendering of the web interface are based on HTML and CSS. Multiple statistical chart formats for displaying query results were implemented by means of Ag-Grid and Chart.js. JBrowse2^[30] was put on the web server so that people could look at the genomic track information for all ac4C modification sites or SNPs. Also, RNAfold was deployed on the server for real time calculation and visualization of RNA secondary structure and minimum free energy (MFE). The previously built multi-model framework for ac4C site prediction, as well as custom python and r scripts, have been deployed on the server, which allows for real time prediction of human ac4C modification sites and interactive analysis of how genetic variants affect ac4C modifications.

3. Results

3.1. ac4C sites in ac4CDB

ac4CDB database contains 168 experimentally identified ac4C sites from NGS studies(Supplementary Table 1)^[31–55], across 9 species: human (1170056), mouse (418441), rat (3169), zebrafish (12848), worm (5984), dog(104653), Arabidopsis thaliana (24051), rice (94119), and tomato (14259). Humans have ac4C sites divided into two groups according to how they were detected: single base resolution (1004 sites, ac4C-seq) and ac4C enriched regions (1846656 sites, acRIP-seq); for every other species, only acRIP-seq was used to find these places. Besides being technically classified, the database also systematically arranges ac4C sites based on their tissue type or cell line. Human sites cover different cellular contexts like HONE (80908 sites), HK-2 (30350), endometrium (71757), MOLM-13 (113938), FLS (9542), SW620 (20743), PANC-1 (20329), heart (49109), UM-UC-3 (44110), CD4 + T cells(6152), DLD-1(151900), SW480(70653), HN6&FaDu(42873), 143B(17844), hESCs(70262), AGS(54460), iSLK(31815), HeLa(48444), MM1.S(61237), and A549(122258). Mouse sites: hippocampus (102191), Hepa1-6 (48962), PFC (47487), uterus (47417), BMDMs (67340), and CD3 + T cells (91060). And plants too like

tomatoes with tissue specific annotation of leaves and seedlings. Collectively, ac4CDB is currently the most complete resource for ac4C acetylation available, with the largest number of species, tissues, and cell lines included (Table 2).

3.2. The ac4C influence on the pathogenesis of diseases involved in ac4CDiseaseDB

ac4CDiseaseDB contains a total of 8943151 genetic variants that may alter ac4C acetylation status in humans, offering the most comprehensive landscape to date of genetic factors potentially linked to ac4C dysregulation. In order to find out how the disease phenotypes work on the epitranscriptomic level, we mapped all the collected human ac4C-related variants onto pathogenic TagSNPs. We found 355365 ac4C-related variants from all genes, which had 352729 known disease phenotypes. And 161714 of those variants came from TCGA somatic mutations in cancers, pointing toward ac4C acetylation's possible part in 33 kinds of human cancer.

3.3. The Web interface and usage

The overall architecture of the ac4C Info platform consists of four functional modules (Fig. 1). The Web interface of ac4C Info is freely accessible, providing a rich, fast and user-friendly one-stop knowledge base for ac4C research. It enables users to quickly query, conduct customized searches, and freely download all collected datasets. Four main modules are presented in ac4C Info, namely ac4CDB, ac4CDiseaseDB, ac4C Predictor and ac4C SNP IFer.

3.4. ac4CDB

The ac4CDB module contains ac4C sites that have been validated through high-throughput sequencing methods. Users can explore and visualize the characteristics of ac4C modifications across different species using various analytical approaches (Fig. 2A). For example, by clicking the ac4CDB button in the top menu bar, users can access the modification site database page. The resulting page initially presents a statistical overview of the collected ac4C sites, categorized by species, analytical techniques, gene types, and other relevant attributes (Fig. 2B). Below this summary there are several drop down menus which allow for additional filtering based on specific criteria of interest-like Tissue/Cell Line (Fig. 2C).

In each header of the data table there are three positional toolbars. First toolbar lets you pin chosen columns to either left or right side of the table and resize them for better readability. Second toolbar has filter and search functions so that people could get the data related to certain genomic region of interest (Fig. 2D). When we use our own filters, the Web will immediately update itself with only the filtered info. Users can also click on the "Detail" link next to any site if they want more details on a particular ac4C site such as the RNA secondary structure predicted by RNAfold, possible miRNA targets, and RBP binding sites. Or they can click the "JBrowse" button for graphical visualizations of the genomic context around the site via an integrated genome browser (Figs. 2E and 2F). Third toolbar is used to toggle the visibility of individual columns so that user can choose which information to show based on his requirement.

3.5. ac4CDiseaseDB

ac4CDiseaseDB (Fig. 3A) curates experimentally supported variations and their disease associations related to acetylated cytidine (ac4C). Users can query individual ac4C-related single nucleotide polymorphisms (SNPs) along with comprehensive annotations, including reference and mutant sequences, the relative genomic position of each SNP, predicted miRNA targets, RNA-binding protein (RBP) interactions, cross-references to public databases such as dbSNP, 1000 Genomes(Phase 3), and The Cancer Genome Atlas (TCGA), as well as epigenomic impacts on ac4C status—specifically, gain or loss of ac4C function(Fig. 3B). by utilizing the 'ClinVarNumber' (Fig. 3C) column in the data table header for rapid sorting, users can prioritize entries based on their clinical relevance. Furthermore, the search box located in the header section enables queries for all ac4C-associated variants linked to specific disease phenotypes, in addition to alternative search criteria such as gene symbols and genomic coordinates. Consistent with ac4CDB, ac4CDiseaseDB also features a "Detail" button that provides in-depth information for each entry, including TagSNPs mapped to the site from ClinVar data (Fig. 3D).

3.6. Analysis modules (ac4C Predictor and ac4C SNP IFer) and auxiliary tools and downloads

In order to do interactive analysis on the data we collected, we created two new analytical modules based off of nine pre-existing deep neural network models. ac4C predictor allows you to predict where there might be ac4c sites given an RNA sequence that someone submitted in FASTA format. Required input sequence length depends on the parameters of the underlying prediction model (Fig. 4A). Users can freely download the predictions made by ac4C predictor and also get detailed annotations for all the columns in the output (Fig. 4B). Besides ac4C Predictor there is also ac4C SNP IFer module which enables user to determine the possible effect of SNPs of interest on ac4C epitranscriptome in a particular species. ac4C SNP IFer takes a standard VCF file with genetic variants as input and calculates the difference in ac4C association scores between reference and mutant RNA sequences. And we can get those output files for free, they have some information about every column (Fig. 4C and 4D).

In order to help users prepare their inputs, we've created an additional tool named Sequence Cutter that enables users to shorten sequences based on the exact length requirement for prediction models and also gives them the choice to have the cut centered around cytosine (C) residues. And also, ac4C Info provides downloading for all curated data sets. In the "Download" section, you can choose and get any dataset in compressed form.

4. Discussion

Epitranscriptomics is a field of study after the genomic age. Among these, RNA modification, which is an important epigenetic mechanism that regulates gene expression, takes part in many biological processes such as RNA splicing, transportation, stability, and translation efficiency, greatly affecting basic life activities like cell growth, maturation, and death. The abnormality of RNA modification is

associated with the onset and progression of different kinds of illness. In all the kinds of RNA modifications, N4-acetylcytidine(ac4C) is a kind of conserved acetylation modification which can be found in both prokaryotic and eukaryotic RNA molecules such as mRNAs, tRNAs and rRNAs, it becomes one of the hotspots in the field of epitranscriptomics.

High throughput sequencing technology like NGS and ONT with its fast growing and wide application have provided a large amount of sequencing sample which is an important basis for the full analysis of ac4C(N4-acetylcytidine) RNA modification under different biological circumstances. ac4C is also a kind of RNA modification that's very important because it gets involved in all sorts of big things like controlling how long mRNAs last, making sure proteins get made faster, helping RNA fold up properly, and letting cells grow and change. Abnormal regulation of ac4C is associated with tumor, neurodegenerative disease, and other pathologies, making it a hot topic in the field of epitranscriptomics. In this situation, building a one-stop ac4C data integration and analysis platform is very important for comprehending the biological function, molecular mechanism, and clinical application value of ac4C modification.

Currently there is some ac4C related databases which gives the researcher places to store and annotate ac4C modification site. But compared to ac4C Info, current ac4C databases are still lacking in many aspects. In terms of data coverage, the ac4C sites contained in the existing database are mostly based on the identification results of samples obtained through traditional sequencing technology, without systematic collection of ac4C sites predicted from RNA sequencing samples. The breakthrough of ONT technology in directly detecting RNA modifications can catch modification-related distinctive signals, giving a fresh way to precisely find ac4C sites. But these newly emerged data haven't been integrated enough, so there are quite a few gaps in both quantity and kind of ac4C sites in present databases, which makes it hard to give an overall picture of what ac4C modification looks like using different sequencing methods.

Secondly, in the ac4C Info, at the level of association analysis of ac4C-related variations, the existing resources have extremely limited collection and annotation of variations that affect ac4C modification (especially disease-related SNPs). SNPs in the ac4C modification sites and their surrounding regions may affect the ac4C modification status by changing the binding efficiency of modification enzymes, RNA secondary structure, etc., thereby regulating the expression of target genes and participating in the occurrence and development of diseases. However, there is currently no database system that collects ac4C-related disease SNPs, and there is also a lack of association annotations with specific disease phenotypes, which greatly limits the in-depth exploration of the role of ac4C modification in disease pathogenesis. In addition, the existing ac4C databases have insufficient analytical tool support and lack efficient analysis modules based on deep learning, making it difficult to meet the precise mining needs of researchers for the inherent data in the database and the data they upload independently, such as the rapid prediction of ac4C sites, quantitative analysis of modification levels, and screening of differentially modified sites.

It is worth noting that the current analysis of ac4C-modified landscapes still faces significant technical and data limitations. On the one hand, the identification of presumed ac4C modifications in direct RNA sequencing samples mostly relies on deep learning model predictions based on the characteristics of modified nucleotides. Because of the complexity of ac4C modification signal and interference from different sample background, the prediction result's reliability still needs to be enhanced. Though we have the chance through direct RNA sequencing to see the ac4C mod, there isn't any good software that can find it well in the raw data, so it makes things slow and hard to get right when we try to find it. On the contrary, there are big differences in how many ac4C sequencing samples there are between different kinds of animals. Existing databases mainly collect ac4C sites from model organisms such as humans and mice, but for species like plants and fungi that have very few sequencing samples, the collected sites do not reflect the overall distribution characteristics of ac4C modifications in these species, limiting comparisons of ac4C modifications across evolution.

In order to solve these problems, the future improvement and development of the ac4C Info database will be carried out according to the following main directions. First, promote the overall upgrade of the database, systematically incorporate ac4C data generated by various technologies like NGS and ONT direct RNA sequencing, especially add the collection of putative ac4C sites in ONT samples, greatly increase data coverage and diversity, and present the ac4C modification panorama in a multi-dimensional manner. Second, create an ac4C-related variant database module which focuses on gathering and labeling disease-related SNPs that impact ac4C modification and makes clear how they relate to different disease appearances, giving important sources for knowing what happens when there is abnormal ac4C modification. Third, make better deep learning based special analysis tools so we can do fast analysis and dig deep into the information in the database and stuff people upload. Fourth, set up a routine update system, regularly include the newest technologies and data to directly recognize ac4C modification as direct RNA sequencing technology keeps maturing; meanwhile, particularly add sequencing samples and ac4C site data of rare species, gradually improve the ac4C modification maps of various species.

To sum up, ac4C modification is an important part of the study area of epitranscriptomics, and integrating and improving the related data sources has great scientific value. To establish a full, fast, and constantly updated one-stop shop for ac4C, so that it will give top-notch information and analysis help to the worldwide ac4C research group, to push forward thorough investigations into ac4C's biological roles, molecular workings, and medical uses, and to assist with the discovery of what epitranscriptomic regulation does at the center of health and sickness.

5. Conclusion

We have constructed the ac4C Info platform through a unified bioinformatics pipeline and web page technology. This platform includes a high-quality ac4C site database, ac4CDB, covering nine species including human and mouse; an extended database, ac4CDiseaseDB, exploring the potential role of ac4C acetylation in disease pathogenesis; and tools for predicting RNA ac4C acetylation modification sites

and analyzing the impact of genetic variations on ac4C modification. In summary, these several functional modules have jointly built the most comprehensive, species-wide, detailed-annotated, and method-rich ac4C RNA acetylation modification database to date.

Declarations

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

All the result data of this study can be freely downloaded from the download area of the database platform at: <https://hebaubioinformatics.cn:447/download.php>. The used codes and scripts can be obtained from the corresponding author upon reasonable request. Supplementary materials can be found in the online DOI.

References

1. Liu, R., Wubulikasimu, Z., Cai, R., et al.: NAT10-mediated N4-acetylcytidine mRNA modification regulates self-renewal in human embryonic stem cells [J]. *Nucleic Acids Res.* **51**(16), 8514–8531 (2023)
2. Xu, T., Wang, J., Wu, Y., et al.: Ac4C Enhances the Translation Efficiency of Vegfa mRNA and Mediates Central Sensitization in Spinal Dorsal Horn in Neuropathic Pain [J]. *Adv. Sci. (Weinh)*. **10**(35), e2303113 (2023)
3. Gu, Z., Zou, L., Pan, X., et al.: The role and mechanism of NAT10-mediated ac4C modification in tumor development and progression [J]. *MedComm* 2024, 5(12): e70026. (2020)
4. Zhao, J., Liu, H., Li, T., et al.: N4-acetylcytidine mRNA modification regulates photosynthesis in plants [J]. *New. Phytol.* **247**(5), 2098–2117 (2025)
5. Liang, Z., Ye, H., Ma, J., et al.: m6A-Atlas v2.0: updated resources for unraveling the N6-methyladenosine (m6A) epitranscriptome among multiple species [J]. *Nucleic Acids Res.* **52**(D1), D194–D202 (2024)
6. Xuan, J., Chen, L., Chen, Z., et al.: RMBase v3.0: decode the landscape, mechanisms and functions of RNA modifications [J]. *Nucleic Acids Res.* **52**(D1), D273–D284 (2024)
7. Ma, J., Song, B., Wei, Z., et al.: m5C-Atlas: a comprehensive database for decoding and annotating the 5-methylcytosine (m5C) epitranscriptome [J]. *Nucleic Acids Res.* **50**(D1), D196–D203 (2022)
8. Wang, X., Zhang, Y., Chen, K., et al.: m7GHub V2.0: an updated database for decoding the N7-methylguanosine (m7G) epitranscriptome [J]. *Nucleic Acids Res.* **52**(D1), D203–D212 (2024)

9. Members, C.-N., Partners: Nucleic Acids Res. **53**(D1), D30–D44 (2025). Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2025 [J]
10. Chen, S., Zhou, Y., Chen, Y., et al.: fastp: an ultra-fast all-in-one FASTQ preprocessor [J]. Bioinformatics. **34**(17), i884–i890 (2018)
11. Kim, D., Paggi, J.M., Park, C., et al.: Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype [J]. Nat. Biotechnol. **37**(8), 907–915 (2019)
12. Li, H., Handsaker, B., Wysoker, A., et al.: The Sequence Alignment/Map format and SAMtools [J]. Bioinformatics. **25**(16), 2078–2079 (2009)
13. Meng, J., Cui, X., Rao, M.K., et al.: Bioinformatics. **29**(12), 1565–1567 (2013). Exome-based analysis for RNA epigenome sequencing data [J]
14. Quinlan, A.R., Hall, I.M.: BEDTools: a flexible suite of utilities for comparing genomic features [J]. Bioinformatics. **26**(6), 841–842 (2010)
15. Zhu, Y., Xu, G., Yang, Y.T., et al.: POSTAR2: deciphering the post-transcriptional regulatory logics [J]. Nucleic Acids Res. **47**(D1), D203–D211 (2019)
16. Li, J.H., Liu, S., Zhou, H., et al.: starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data [J]. Nucleic Acids Res. **42**(Database issue), D92–97 (2014)
17. Lai, F.L., Gao, F.: LSA-ac4C: A hybrid neural network incorporating double-layer LSTM and self-attention mechanism for the prediction of N4-acetylcytidine sites in human mRNA [J]. Int. J. Biol. Macromol. **253**(Pt 3), 126837 (2023)
18. Liu, R., Zhang, Y., Wang, Q., et al.: TransAC4C-a novel interpretable architecture for multi-species identification of N4-acetylcytidine sites in RNA with single-base resolution [J]. Brief. Bioinform, **25**(3). (2024)
19. Lu, R., Qiao, J., Li, K., et al.: ERNIE-ac4C: A Novel Deep Learning Model for Effectively Predicting N4-acetylcytidine Sites [J]. J. Mol. Biol. **437**(6), 168978 (2025)
20. Jia, Y., Zhang, Z., Yan, S., et al.: Voting-ac4C:Pre-trained large RNA language model enhances RNA N4-acetylcytidine site prediction [J]. Int. J. Biol. Macromol. **282**(Pt 3), 136940 (2024)
21. Jia, J., Cao, X., Wei, Z.: DLC-ac4C: A Prediction Model for N4-acetylcytidine Sites in Human mRNA Based on DenseNet and Bidirectional LSTM Methods [J]. Curr. Genomics. **24**(3), 171–186 (2023)
22. Wang, C., Ju, Y., Zou, Q., et al.: DeepAc4C: a convolutional neural network model with hybrid features composed of physicochemical patterns and distributed representation information for identification of N4-acetylcytidine in mRNA [J]. Bioinformatics. **38**(1), 52–57 (2021)
23. He, W., Han, Y., Zuo, Y., et al.: NBCR-ac4C: A Deep Learning Framework Based on Multivariate BERT for Human mRNA N4-Acetylcytidine Sites Prediction [J]. J. Chem. Inf. Model. **64**(20), 8074–8081 (2024)
24. Yi, M., Zhou, F., Deng, Y.: STM-ac4C: a hybrid model for identification of N4-acetylcytidine (ac4C) in human mRNA based on selective kernel convolution, temporal convolutional network, and multi-

- head self-attention [J]. *Front. Genet.* **15**, 1408688 (2024)
25. Zhang, S., Liu, K., Xu, Y., TransCNN: A novel architecture combining transformer and TextCNN for detecting N4-acetylcytidine sites in human mRNA [J]. *Anal. Biochem.* **703**, 115882 (2025)
 26. Sherry, S.T., Ward, M.H., Kholodov, M., et al.: dbSNP: the NCBI database of genetic variation [J]. *Nucleic Acids Res.* **29**(1), 308–311 (2001)
 27. Dyer, S.C., Austine-Orimoloye, O., Azov, A.G., et al.: Ensembl 2025 [J]. *Nucleic Acids Res.* **53**(D1), D948–D957 (2025)
 28. Tomczak, K., Czerwińska, P., Wiznerowicz, M.: The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge [J]. *Contemp. Oncol. (Pozn).* **19**(1A), A68–77 (2015)
 29. Landrum, M.J., Lee, J.M., Benson, M., et al.: ClinVar: improving access to variant interpretations and supporting evidence [J]. *Nucleic Acids Res.* **46**(D1), D1062–D1067 (2018)
 30. Diesh, C., Stevens, G.J., Xie, P., et al.: JBrowse 2: a modular genome browser with views of syntenic and structural variation [J]. *Genome Biol.* **24**(1), 74 (2023)
 31. Guo, G., Shi, X., Wang, H., et al.: Epitranscriptomic N4-Acetylcytidine Profiling in CD4(+) T Cells of Systemic Lupus Erythematosus [J]. *Front. Cell. Dev. Biol.* **8**, 842 (2020)
 32. Xiang, Y., Zhou, C., Zeng, Y., et al.: NAT10-Mediated N4-Acetylcytidine of RNA Contributes to Post-transcriptional Regulation of Mouse Oocyte Maturation in vitro [J]. *Front. Cell. Dev. Biol.* **9**, 704341 (2021)
 33. Wei, R., Cui, X., Min, J., et al.: NAT10 promotes cell proliferation by acetylating CEP170 mRNA to enhance translation efficiency in multiple myeloma [J]. *Acta Pharm. Sin. B.* **12**(8), 3313–3325 (2022)
 34. Jin, C., Wang, T., Zhang, D., et al.: Acetyltransferase NAT10 regulates the Wnt/ β -catenin signaling pathway to promote colorectal cancer progression via ac(4)C acetylation of KIF23 mRNA [J]. *J. Exp. Clin. Cancer Res.* **41**(1), 345 (2022)
 35. Deng, M., Zhang, L., Zheng, W., et al.: Helicobacter pylori-induced NAT10 stabilizes MDM2 mRNA via RNA acetylation to facilitate gastric cancer progression [J]. *J. Exp. Clin. Cancer Res.* **42**(1), 9 (2023)
 36. Xie, R., Cheng, L., Huang, M., et al.: NAT10 Drives Cisplatin Chemoresistance by Enhancing ac4C-Associated DNA Repair in Bladder Cancer [J]. *Cancer Res.* **83**(10), 1666–1683 (2023)
 37. Xu, L., Zheng, S., Liu, B., et al.: Epitranscriptomic profiling of N4-acetylcytidine-related RNA acetylation in the spinal dorsal horn of rat with cancer-induced bone pain [J]. *Mol. Pain.* **19**, 17448069231178487 (2023)
 38. Yan, Q., Zhou, J., Wang, Z., et al.: NAT10-dependent N(4)-acetylcytidine modification mediates PAN RNA stability, KSHV reactivation, and IFI16-related inflammasome activation [J]. *Nat. Commun.* **14**(1), 6327 (2023)
 39. Liu, Y., Wang, X., Liu, Y., et al.: N4-acetylcytidine-dependent GLMP mRNA stabilization by NAT10 promotes head and neck squamous cell carcinoma metastasis and remodels tumor microenvironment through MAPK/ERK signaling pathway [J]. *Cell. Death Dis.* **14**(11), 712 (2023)

40. Shi, J., Yang, C., Zhang, J., et al.: NAT10 Is Involved in Cardiac Remodeling Through ac4C-Mediated Transcriptomic Regulation [J]. *Circ. Res.* **133**(12), 989–1002 (2023)
41. Hu, Z., Lu, Y., Cao, J., et al.: N-acetyltransferase NAT10 controls cell fates via connecting mRNA cytidine acetylation to chromatin signaling [J]. *Sci. Adv.* **10**(2), eadh9871 (2024)
42. Zhang, H., Lu, R., Huang, J., et al.: N4-acetylcytidine modifies primary microRNAs for processing in cancer cells [J]. *Cell. Mol. Life Sci.* **81**(1), 73 (2024)
43. Gong, S., Qiao, H., Wang, J.Y., et al.: Ac4C modification of lncRNA SIMALR promotes nasopharyngeal carcinoma progression through activating eEF1A2 to facilitate ITGB4/ITGA6 translation [J]. *Oncogene.* **43**(38), 2868–2884 (2024)
44. Zou, Y., Guo, S., Wen, L., et al.: Targeting NAT10 inhibits osteosarcoma progression via ATF4/ASNS-mediated asparagine biosynthesis [J]. *Cell. Rep. Med.* **5**(9), 101728 (2024)
45. Yang, Y., Lu, J., Liu, Y., et al.: Improvement of MASLD and MASH by suppression of hepatic N-acetyltransferase 10 [J]. *Mol. Metab.* **89**, 102030 (2024)
46. Xue, L., Zhao, J., Liu, X., et al.: MK-801-exposure induces increased translation efficiency and mRNA hyperacetylation of Grin2a in the mouse prefrontal cortex [J]. *Epigenetics.* **19**(1), 2417158 (2024)
47. Zhang, S., Huang, F., Wang, Y., et al.: NAT10-mediated mRNA N(4)-acetylcytidine reprograms serine metabolism to drive leukaemogenesis and stemness in acute myeloid leukaemia [J]. *Nat. Cell. Biol.* **26**(12), 2168–2182 (2024)
48. Zheng, Y., Song, D., Guo, M., et al.: EGFR inhibition augments the therapeutic efficacy of the NAT10 inhibitor Remodelin in Colorectal cancer [J]. *J. Exp. Clin. Cancer Res.* **44**(1), 37 (2025)
49. Sun, L., Li, X., Xu, F., et al.: A critical role of N(4)-acetylation of cytidine in mRNA by NAT10 in T cell expansion and antiviral immunity [J]. *Nat. Immunol.* **26**(4), 619–634 (2025)
50. Cai, L., Li, B., Zhou, Q., et al.: N-acetyltransferase 10 catalyzes RNA N(4)-acetylcytidine to regulate photosynthesis in rice [J]. *Cell. Rep.* **44**(3), 115428 (2025)
51. Huang, L., Lu, Y., He, R., et al.: N(4)-acetylcytidine modification of ITGB5 mRNA mediated by NAT10 promotes perineural invasion in pancreatic ductal adenocarcinoma [J]. *J. Exp. Clin. Cancer Res.* **44**(1), 103 (2025)
52. Zhang, X., Qin, S., Huang, F., et al.: N4-acetylcytidine coordinates with NP1 and CPSF5 to facilitate alternative RNA processing during the replication of minute virus of canines [J]. *Nucleic Acids Res.*, **53**(6). (2025)
53. Ji, H.N., Zhou, H.Q., Qie, J.B., et al.: Dysregulated ac4C modification of mRNA in a mouse model of early-stage Alzheimer's disease [J]. *Cell. Biosci.* **15**(1), 45 (2025)
54. Feng, G., Yin, G., Liu, Y., et al.: Nat10 maintains the homeostasis of pluripotent and 2-cell-like states in mouse embryonic stem cells through mRNA cytidine acetylation [J]. *Nucleic Acids Res.*, **53**(11). (2025)
55. Ma, A., Wang, S., He, X., et al.: Synergistic function of RNA modifications in Arabidopsis and rice [J]. *aBIOTECH.* **6**(4), 803–815 (2025)

Figures

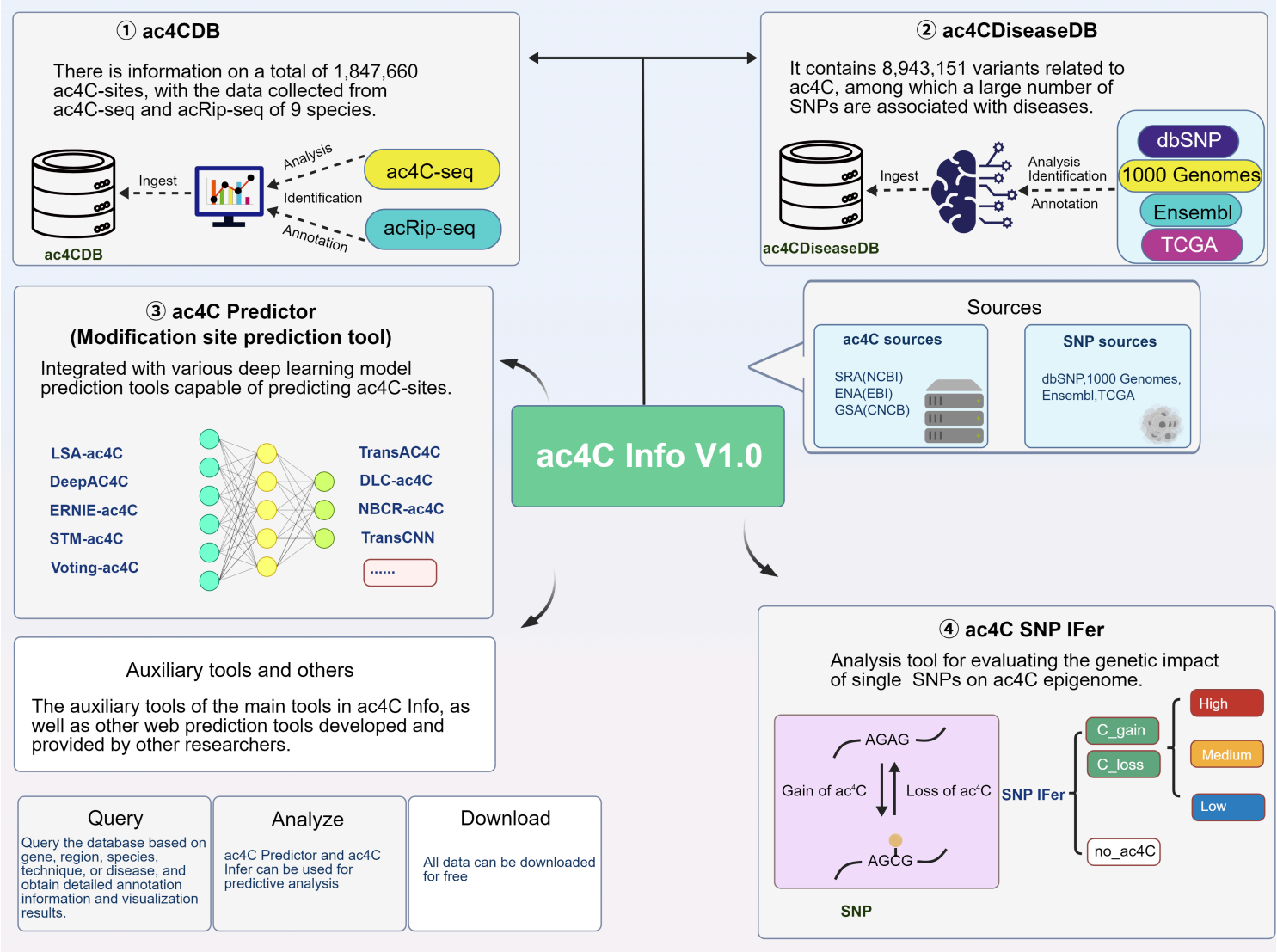


Figure 1

The overall architecture of ac4C Info consists of four main components: (1) ac4CDB: It contains 1,847,660 ac4C sites identified from multiple cell types or cell lines and tissues of 9 species. (2) ac4CDiseaseDB: It includes 8,943,151 pieces of information on functional variations related to ac4C modification. (3) ac4C Predictor: A tool for predicting ac4C sites that incorporates 9 prediction models. (4) ac4C SNP IFer: A tool for assessing how gene variations affect ac4C modification. An integrated web interface provides query, search, visualization, and download functions for all collected data and is freely accessible at: <https://hebaubioinformatics.cn:447/>.

focus on the second column, which provides various filtering conditions for data types or string types. (E) Comprehensive information page of each site. (F) Each site provides visualization through JBrowse2.

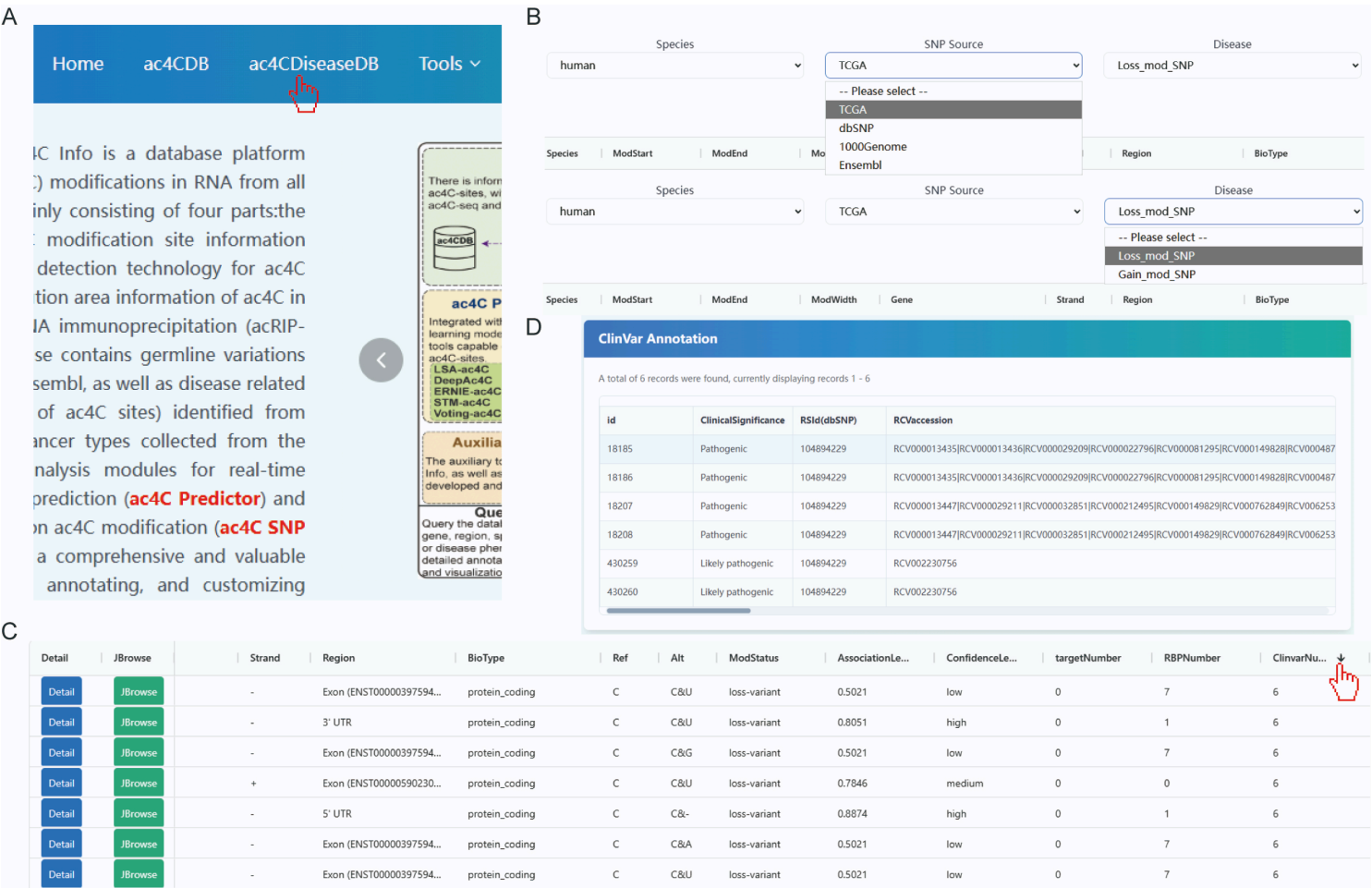


Figure 3

Main functions and contents of ac4CDiseaseDB. (A and B) Show the jump paths of this database and provide several options for further filtering the dataset, including species, SNP Source and SNP type (presumed Disease). (C and D) The data table of ac4CDiseaseDB is associated with Clinvar-related information. The usage method of ac4CDiseaseDB is basically the same as that of ac4CDB.

- [Supplementary.xlsx](#)
- [SupplementaryTable2.xlsx](#)
- [SupplementaryTable1.xlsx](#)
- [SupplementaryTable2.xlsx](#)