



RiboChat: a chat-style web interface for analysis and annotation of ribosome profiling data

Mingzhe Xie[†], Ludong Yang[†], Gennong Chen[†], Yan Wang, Zhi Xie  and Hongwei Wang 

Corresponding authors: Hongwei Wang, State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangzhou 510060, China. Tel: +862066677086; E-mail: biocwhw@126.com; Zhi Xie, State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Visual Science, Guangzhou 510060, China. Tel: +862066677086; E-mail: xiezhi@gmail.com

[†]The first three authors contributed equally to the work.

Abstract

The increasing volume of ribosome profiling (Ribo-seq) data, computational complexity of its data processing and operational handicap of related analytical procedures present a daunting set of informatics challenges. These impose a substantial barrier to researchers particularly with no or limited bioinformatics expertise in analyzing and decoding translation information from Ribo-seq data, thus driving the need for a new research paradigm for data computation and information extraction. In this knowledge base, we herein present a novel interactive web platform, RiboChat (<https://db.cngb.org/ribobench/chat.html>), for direct analyzing and annotating Ribo-seq data in the form of a chat conversation. It consists of a user-friendly web interface and a backend cloud-computing service. When typing a data analysis question into the chat window, the object-text detection module will be run to recognize relevant keywords from the input text. Based on the features identified in the input, individual analytics modules are then scored to find the perfect-matching candidate. The corresponding analytics module will be further executed after checking the completion status of the uploading of datasets and configured parameters. Overall, RiboChat represents an important step forward in the emerging direction of next-generation data analytics and will enable the broad research community to conveniently decipher translation information embedded within Ribo-seq data.

Keywords: ribosome profiling, chat interface, interactive mining, data analysis, visualization

Introduction

By pinpointing the precise positions of ribosomes actively engaged in translation in a cell at a particular moment, ribosome profiling (Ribo-seq) has laid a foundation for the genome-wide analysis of translation *in vivo* at a sub-codon resolution [1, 2]. The capacity of Ribo-seq for translome analysis has opened up a broad range of biological applications across different species and experimental conditions [3–8], leading to major scientific breakthroughs that have greatly advanced our understanding of the composition, regulation and mechanism of translation [9–15]. All of these applications benefit not only from the awesome power of the Ribo-seq technique but also from the availability of specialized computational tools. Indeed, the unique characteristics of Ribo-seq data have impelled the development of a rich variety of excellent specialized tools, such as RiboseQC [16] and RiboVIEW [17] for quality evaluation, ORFquant [18] and

PRICE [19] for actively translated ORF detection, and Xtail [20] and RiboDiff [21] for differential translation analysis. There are currently dozens of such tools specifically developed for Ribo-seq data that have also been systematically reviewed by several groups, including Wang [22], Calviello [23] and Kiniry *et al.* [24]. Indubitably, these tools provide an unprecedented opportunity to decode translation information and derive meaningful insights from Ribo-seq data.

Although these tools become increasingly essential for Ribo-seq data analysis, getting started with them remains challenging, particularly for researchers with no or limited bioinformatics expertise. The main reasons for this include that (i) they are often implemented in different programming languages, leading to an increased demand for diverse programming skills and (ii) their installation often depends on language-centered running environments, leading to an increased difficulty of

Mingzhe Xie is master's student at State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University.

Ludong Yang is PhD student at State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University.

Gennong Chen is master's student at State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University.

Yan Wang is PhD student at State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University.

Zhi Xie (MD, PhD) is a professor of Bioinformatics at Zhongshan Ophthalmic Center, Sun Yat-sen University. He is interested in applying big data analytics in biology and medicine.

Hongwei Wang (PhD) is an associate professor of Bioinformatics at Zhongshan Ophthalmic Center, Sun Yat-sen University. His research is focused on gene translation and translational regulation.

Received: September 4, 2021. **Revised:** November 29, 2021. **Accepted:** December 8, 2021

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

software configuration management. All of these factors substantially increase the time spent in data processing activities. Thus, there is an urgent demand for web-based analytics tools that enable users to perform various analyses or the visualization of Ribo-seq data without extensive programming skills.

Regrettably, this type of analytics tool is still very rarely used. RiboGalaxy [25] is the first web-based platform for Ribo-seq data analysis with visualization functionality, but its significant components, such as workflow and visualization, are only available after login, and the many options of the website will significantly reduce the overall appeal of its usability. Although Riboviz and RiboToolkit provide user-friendly web applications for the exploration and analysis of Ribo-seq data online, the former only allows the viewing of pre-executed analyses of preloaded datasets [26], whereas the latter only supports a predefined set of analyses [27], in contrast to RiboGalaxy, which allows the flexible customization of analytical pipelines. Despite these impressive achievements, it still takes users considerable time and effort to identify appropriate tools and learn distinct web interfaces.

With the gradual deepening of translome research and the continuous accumulation of Ribo-seq data [28, 29], analyzing and decoding translation information from Ribo-seq data require revolutionized paradigm of data analysis. As pointed out in a pioneering paper [30], next-generation data analytics should possess key features such as understanding natural language, artificial intelligence, transparency, friendliness and crowdsourcing. Following this premise, we developed RiboChat for direct Ribo-seq data analysis in a chat conversation manner. Through a chat conversation with RiboChat, the user can easily complete the entire process with detailed step-by-step guides, from raw read filtering, trimming, and alignment to quality evaluation, footprint meta-gene analysis, and expression quantification and, finally, actively translated ORF detection, differential translation analysis, and analytical result visualization, along with automated report generation.

Architectural overview

The overall architecture of RiboChat is outlined in Figure 1. It mainly consists of a frontend web interface and a backend cloud-computing service. The frontend provides an online chat-style web interface, including an input area for chatting and an output area where the chat conversation is written. All conversational interactions with users are based on human languages. The backend includes four major parts: object-text detection module, question answering module, data analytics module and feedback module. The object-text detection module is used to recognize relevant keywords from the input text, such as species (e.g. human or mouse) and data type (e.g. ribo-seq or RNA-seq). The question-answering module evaluates the similarity between the presented question and each candidate answer to find the best-matching

answer. During question answering, some key parameters are recorded to create a specific analytical task. The recording analytical task then calls the corresponding data analytics module and, thus, executes the entire data analysis after confirming that the detected task is indeed the intended analysis. When the analysis is completed, the report generator is initiated to generate HTML code for a single webpage that summarizes all analysis results.

Analytics module summary

As an initial effort, we built nine data analytics modules covering the most popular steps of Ribo-seq analysis, including quality checks, filtering and trimming, alignment, counting and normalization of the sequenced reads and, very often, postmapping quality inspection, ORF detection, differential translation analysis, functional enrichment analysis and integrative exploration of analytical results (Figure 1). When receiving FASTQ files of Ribo-seq data, an initial check of sequence quality based on diagnostic quality plots is essential. The outstanding tools FastQC [31] and MultiQC [32] are preinstalled to assess the overall quality of the sequence reads. To further increase the quality and reliability of the analysis, the trimming of adaptor sequences and low-quality bases together with the widely used method of filtering read lengths are very important components of the data analysis. These processes are achieved by the ultrafast all-in-one FASTQ preprocessor fastp [33]. Typically, the next step is the alignment of reads to a reference genome, which is conducted with two fast, sensitive aligners, HISAT2 [34] and STAR [35]. Notably, to save personnel time and simplify operations, reference genomes (FASTA format) and gene annotation files (GTF format) for multiple species (including eukaryotes—*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Danio rerio*, *Drosophila melanogaster*, *Gallus gallus*, *Homo sapiens*, *Mus musculus*, *Rattus norvegicus* and *Saccharomyces cerevisiae*; and prokaryotes—*Bacillus subtilis*, *Escherichia coli* strain K12, *Escherichia coli* strain Sakai, *Halobacterium salinarum*, *Pseudomonas fluorescens*, *Salmonella enterica* and *Streptomyces albidoflavus*) are built into this module. Once the reads have been mapped, they can generally be assigned to a gene or a transcript with higher confidence, which is known as counting. Here, a general-purpose read summarization program, featureCounts [36], is used for the inference of gene and isoform abundance. After counting, normalization is of vital importance to accurately interpret the results of experiments since sample-specific systematic biases, along with distortions such as those affecting the overall distribution of count data, can introduce unwanted variation in Ribo-seq data that will obscure the underlying biological signal. Multiple normalization strategies are presented here, including the reads per kilobase per million mapped reads, fragments per kilobase per million mapped reads and transcripts per million methods.

Because of the unique characteristics of Ribo-seq data, one critical step is the evaluation of postmapping

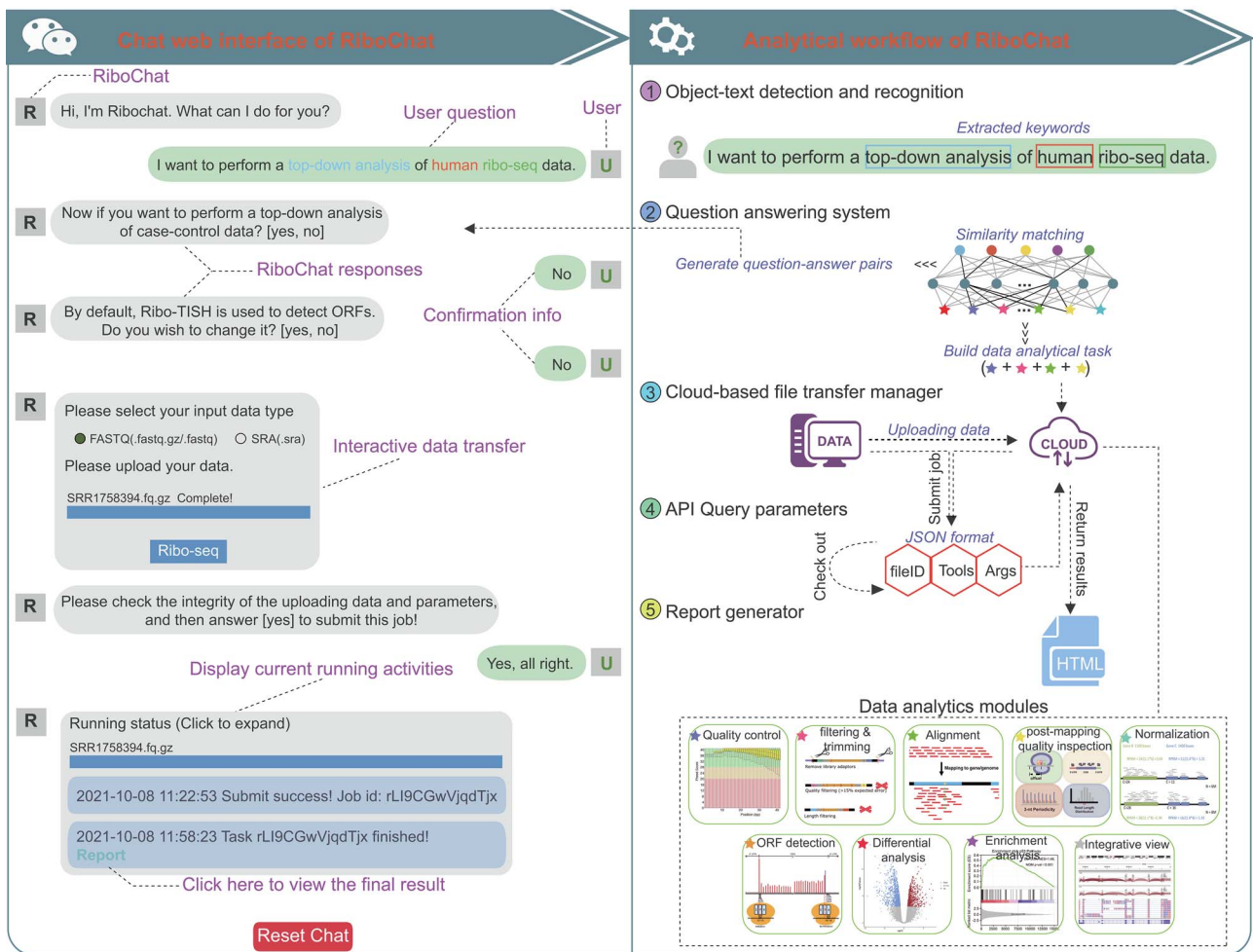


Figure 1. An overview of the chat interface and analytical workflow of RiboChat.

quality, including the assessment of the read-length distribution, codon-frame phasing, triplet periodicity and read counts across genomic features (e.g. coding sequence, untranslated regions, noncoding RNAs and mitochondria), and often, the computation of various metagene plots. RiboseQC is responsible for confirming these characteristic features of high-quality Ribo-seq libraries and is already installed [16]. Such strategies have opened avenues for re-evaluating the coding capacity of the genome, particularly in many genomic regions outside of annotated protein-coding genes that were previously assumed to be noncoding regions [37]. Several state-of-the-art tools, including Ribo-TISH [38], PRICE [19], RiboCode [39] and ORFquant [18], are preinstalled, providing alternative options to cater to the demands of different user preferences. For most Ribo-seq studies, the data analyses also include the following key steps: differential translation efficiency analysis and functional enrichment analysis. To facilitate differential analysis, DESeq2 [40] and Xtail [20] are preinstalled to identify quantitative changes in translation between experimental groups and even in translational efficiency levels when uploading match RNA-seq and Ribo-seq data. To facilitate enrichment analysis, clusterProfiler

[41] is preinstalled to unveil biological functions and pathways of differential genes. In addition, to gain an integrative view of these results, the IGV-Web application [42] is already installed for interactive exploration of the analytical results through a web browser.

It should be noted that some initialization parameters can also be modified, including those related to (i) adapter contents in sequences, (ii) minimum Phred quality scores for base calling, (iii) the minimum length requirement for a read, (iv) the maximum number of mismatches for a read, (v) the maximum number of multiple alignments for a read, (vi) the selection of a typical length range of footprints and (vii) cut-off thresholds for differential translation analysis and functional enrichment analysis. The modular design of these analytical functions enables the implementation of flexible workflows in response to different analytical tasks.

Platform implementation

RiboChat is deployed on the Alibaba Cloud Elastic Compute Service with a 32-core CPU and 256 GB RAM. The frontend web interface is built using HTML5, CSS and JavaScript libraries, including jQuery (<http://jquery.com>), Bootstrap (<http://getbootstrap.com/>) and popper

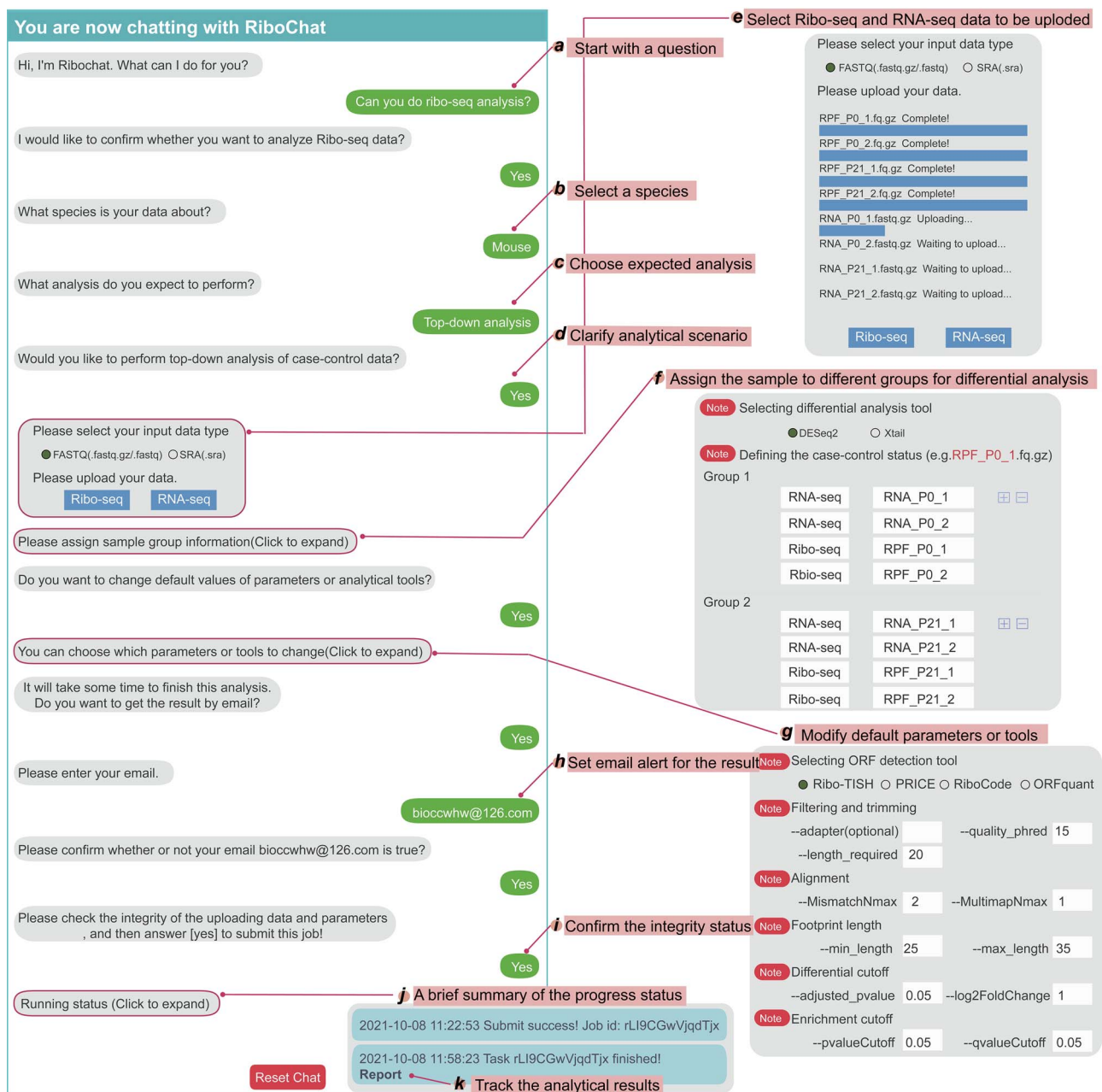


Figure 2. An operational illustration of a top-down analysis of a case-control study.

(<http://popper.js.org>). The backend is implemented by using the PHP framework swoole (<https://www.swoole.com/>) and is mainly responsible for frontend and backend communication, task delivery, tool connection, results integration and so on. A diverse set of tools has been preinstalled and further connected using snakemake [43], which is a workflow management system written in Python. Unique snakemake files are written separately for individual tools and are then automatically tied together according to the frontend transformed parameters.

Example illustration

Given that Ribo-seq and RNA-seq are usually performed in parallel, RNA-seq data can also be uploaded and analyzed. Depending on the types and conditions

of uploading data, there are five scenarios of data analytics that are hereby established: case/control-only experimental design allowing for analyzing uniq-condition Ribo-seq data or uniq-condition RNA-seq data and case-control experimental design allowing for analyzing multi-condition Ribo-seq data, RNA-seq data, or matched Ribo-seq and RNA-seq data. Currently, RiboChat only accepts files in the commonly used text-based formats FASTQ or gz-compressed FASTQ as input. It supports the upload of a single file or several individual files at once. Although there is no limitation in terms of uploaded file numbers or size, we have introduced a queue management system considering the performance of the cloud-computing server in which a maximum of four files are allowed to run in parallel at one time.

One example of a top-down analysis of a case-control study (that is, two-condition matched mouse Ribo-seq and RNA-seq data) is presented here to illustrate how to quickly get started with RiboChat. The detailed chat conversations and operational processes are shown in [Figure 2](#). A user can start this analysis by simply asking, 'Could you do a Ribo-seq analysis?'. After confirming the answer from RiboChat, the next steps are to choose the appropriate species and the expected analysis (herein, 'mouse' and 'top-down analysis', respectively), followed by the clarification of the case-control analytical scenario and default software requirement under the guidance of the chat dialogue. The uploaded data panel will be displayed, and the 'Ribo-seq' and 'RNA-seq' buttons are then clicked to select the corresponding Ribo-seq and RNA-seq data and upload them to the cloud server. A progress bar shows the uploading progress in real time. The text 'Complete!' will appear when all data have been successfully uploaded. This is followed by the assignment of sample group information to perform differential translation analysis. Then, RiboChat will ask, 'Do you want to change default values of parameters or analytical tools?'. In reply to a 'yes' response, the parameters and tools are shown in the expanded panel that can be modified directly. For instance, by default, the maximum number of allowed mismatches is set to tolerate up to two mismatches; the maximum number of alignments allowed for a read is limited to unique mapping; and the minimum Phred quality score for base calling is set to 15. The recommended size range of ribosome footprints is between 25 and 35 nt. The default cut-offs for differential translation analysis are an adjusted P value of less than 0.05 and a fold change greater than 2. Notably, for a dataset of ~50 GB, it will take approximately 1 h to complete the entire data analysis. As the chat continues, RiboChat will ask, 'Do you want to get the result by e-mail?'. After entering an e-mail address and confirming this submission, a unique identifier composed of 15 random characters is assigned to the job, which helps the user quickly track the analytical task. A brief summary of the progress status of the current job will be shown in the chat window. Once it is completed, a result report in HTML format will be generated, and a task notification email will be automatically sent to the user. The user can track the results either directly, by clicking the unique identifier in the progress status report, or by clicking the URL in the task completion notification email. The analytical results are organized in a modular format to enable the user to interactively visualize and explore the results.

Conclusion and discussion

To the best of our knowledge, RiboChat is the first online interactive platform for direct Ribo-seq data analysis implemented via a chat conversation. It enables the broader research community to conveniently and easily perform Ribo-seq data analysis. Compared with existing

tools, the primary added value of RiboChat is that the user interface with human language modalities will enhance human-computer interaction, thereby greatly reducing the barrier to entry for Ribo-seq data analysis, particularly for users with no or limited programming or bioinformatics skills.

As an initial attempt, RiboChat in its present form may be far from perfect, but certainly it represents an important step forward in the emerging direction of next-generation data analytics. In future, to further enhance the efficiency of its applications in the translational research community, we will continue to improve the core object-text detection and question answering modules toward functioning as an intelligent brain to understand and derive precise meaning from human languages. We will also continue to add new data analytics and data visualization modules into RiboChat, making it a truly comprehensive analytics platform to meet different analytical needs. With these improvements and additions, we believe that RiboChat will provide an unprecedented level of convenience for researchers to decode the translation information embedded within Ribo-seq data.

Key points

- RiboChat is a novel interactive web-based platform for direct analyzing and annotating Ribo-seq data in the form of a chat conversation.
- The modular design of its analytical functions enables the implementation of many flexible workflows in response to different analytical tasks.
- RiboChat represents an important step forward in the emerging direction of next-generation data analytics, enabling the broad research community to conveniently decode translation information.

Data Availability

RiboChat is available at <https://db.cngb.org/ribobench/chat.html>. All demo data used in the study are retrieved from the Gene Expression Omnibus data repository under accession number GSE64962 and GSE94982.

Acknowledgments

We would like to thank the support from the China National GeneBank and Center for Precision Medicine, Sun Yat-sen University.

Funding

This work was supported in part by the National Natural Science Foundation of China (31871302 to Z.X.) and

the Joint Research Fund for Overseas Natural Science of China (31829002 to Z.X.).

References

- Ingolia NT, Brar GA, Rouskin S, et al. The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments. *Nat Protoc* 2012;**7**:1534–50.
- Ingolia NT, Ghaemmaghami S, Newman JRS, et al. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 2009;**324**:218–23.
- Stern-Ginossar N, Ingolia NT. Ribosome profiling as a tool to decipher viral complexity. *Annu Rev Virol* 2015;**2**:335–49.
- Reixachs-Solé M, Ruiz-Orera J, Albà MM, et al. Ribosome profiling at isoform level reveals evolutionary conserved impacts of differential splicing on the proteome. *Nat Commun* 2020;**11**:1–12.
- Fremin BJ, Sberro H, Bhatt AS. MetaRibo-Seq measures translation in microbiomes. *Nat Commun* 2020;**11**:3268.
- Finkel Y, Mizrahi O, Nachshon A, et al. The coding capacity of SARS-CoV-2. *Nature* 2021;**589**:125–30.
- Rubio A, Ghosh S, Müllender M, et al. Ribosome profiling reveals ribosome stalling on tryptophan codons and ribosome queuing upon oxidative stress in fission yeast. *Nucleic Acids Res* 2021;**49**:383–99.
- Shalgi R, Hurt JA, Krykbaeva I, et al. Widespread regulation of translation by elongation pausing in heat shock. *Mol Cell* 2013;**49**:439–52.
- Ingolia NT. Ribosome profiling: new views of translation, from single codons to genome scale. *Nat Rev Genet* 2014;**15**:205–13.
- Brar GA, Weissman JS. Ribosome profiling reveals the what, when, where and how of protein synthesis. *Nat Rev Mol Cell Biol* 2015;**16**:651–64.
- Ingolia NT. Ribosome footprint profiling of translation throughout the genome. *Cell* 2016;**165**:22–33.
- Ingolia NT, Hussmann JA, Weissman JS. Ribosome profiling: global views of translation. *Cold Spring Harb Perspect Biol* 2019;**11**:a032698.
- Li H, Xie M, Wang Y, et al. riboCIRC: a comprehensive database of translatable circRNAs. *Genome Biol* 2021;**22**:79.
- Wang H, Wang Y, Xie S, et al. Global and cell-type specific properties of lincRNAs with ribosome occupancy. *Nucleic Acids Res* 2017;**45**:2786–96.
- Andreev DE, O'Connor PBF, Loughran G, et al. Insights into the mechanisms of eukaryotic translation gained with ribosome profiling. *Nucleic Acids Res* 2017;**45**:513–26.
- Calviello L, Sydow D, Harnett D, et al. Ribo-seQC: comprehensive analysis of cytoplasmic and organellar ribosome profiling data. *bioRxiv* 2019. [10.1101/601468](https://doi.org/10.1101/601468).
- Legrand C, Tuorto F. RiboVIEW: a computational framework for visualization, quality control and statistical analysis of ribosome profiling data. *Nucleic Acids Res* 2020;**48**:e7.
- Calviello L, Hirsekorn A, Ohler U. Quantification of translation uncovers the functions of the alternative transcriptome. *Nat Struct Mol Biol* 2020;**27**:717–25.
- Erhard F, Halenius A, Zimmermann C, et al. Improved Ribo-seq enables identification of cryptic translation events. *Nat Methods* 2018;**15**:363–6.
- Xiao Z, Zou Q, Liu Y, et al. Genome-wide assessment of differential translations with ribosome profiling data. *Nat Commun* 2016;**7**:11194.
- Zhong Y, Karaletsos T, Drewe P, et al. RiboDiff: detecting changes of mRNA translation efficiency from ribosome footprints. *Bioinformatics* 2017;**33**:139–41.
- Wang H, Wang Y, Xie Z. Computational resources for ribosome profiling: from database to web server and software. *Brief Bioinform* 2019;**20**:144–55.
- Calviello L, Ohler U. Beyond read-counts: Ribo-seq data analysis to understand the functions of the transcriptome. *Trends Genet* 2017;**33**:728–44.
- Kiniry SJ, Michel AM, Baranov PV. Computational methods for ribosome profiling data analysis. *WIREs RNA* 2020;**11**:e1577.
- Michel AM, Mullan JPA, Velayudhan V, et al. RiboGalaxy: a browser based platform for the alignment, analysis and visualization of ribosome profiling data. *RNA Biol* 2016;**13**:316–9.
- Carja O, Xing T, Wallace EWJ, et al. Riboviz: analysis and visualization of ribosome profiling datasets. *BMC Bioinformatics* 2017;**18**:461.
- Liu Q, Shvarts T, Sliz P, et al. RiboToolkit: an integrated platform for analysis and annotation of ribosome profiling data to decode mRNA translation at codon resolution. *Nucleic Acids Res* 2020;**48**:W218–29.
- Wang H, Yang L, Wang Y, et al. RPFdb v2.0: an updated database for genome-wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res* 2019;**47**:D230–4.
- Xie S-Q, Nie P, Wang Y, et al. RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res* 2016;**44**:D254–8.
- Li J, Chen H, Wang Y, et al. Next-generation analytics for omics data. *Cancer Cell* 2021;**39**:3–6.
- Andrews S. FastQC A Quality control tool for high throughput sequence data. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (24 August 2021, date last accessed).
- Ewels P, Magnusson M, Lundin S, et al. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 2016;**32**:3047–8.
- Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 2018;**34**:i884–90.
- Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 2015;**12**:357–60.
- Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 2013;**29**:15–21.
- Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 2014;**30**:923–30.
- Ruiz-Orera J, Verdaguer-Grau P, Villanueva-Cañas JL, et al. Translation of neutrally evolving peptides provides a basis for de novo gene evolution. *Nat Ecol Evol* 2018;**2**:890–6.
- Zhang P, He D, Xu Y, et al. Genome-wide identification and differential analysis of translational initiation. *Nat Commun* 2017;**8**:1749.
- Xiao Z, Huang R, Xing X, et al. De novo annotation and characterization of the transcriptome with ribosome profiling data. *Nucleic Acids Res* 2018;**46**:e61.
- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550.
- Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (N Y)* 2021;**2**:100141.
- Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;**14**:178.
- Köster J, Rahmann S. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics* 2012;**28**:2520–2.