Published online 30 August 2022 *Nucleic Acids Research, 2023, Vol. 51, Database issue* ***D1129–D1137***
https://doi.org/10.1093/nar/gkac732

Exploration and analysis of R-loop mapping data with *RLBase*

Henry E. Miller[1,2,3,*], Daniel Montemayor[4,5], Janet Li[3,6,7], Simon A. Levy[1,3,8,9], Roshan Pawar[3,10], Stella Hartono[11], Kumar Sharma[4,5], Bess Frost[1,8,9], Frédéric Chedin[11] and Alexander J.R. Bishop[1,2,12,*]

[1]Department of Cell Systems and Anatomy, UT Health San Antonio, San Antonio, TX 78229, USA, [2]Greehey Children's Cancer Research Institute, UT Health San Antonio, San Antonio, TX 78229, USA, [3]Bioinformatics Research Network, Atlanta, GA 30317, USA, [4]Department of Medicine, UT Health San Antonio, San Antonio, TX 78229, USA, [5]Center for Precision Medicine, UT Health San Antonio, San Antonio, TX 78229, USA, [6]Bioinformatics Graduate Program, University of British Columbia, Vancouver, BC V6T 1Z2, Canada, [7]Canada's Michael Smith Genome Sciences Center, BC Cancer Research, Vancouver, BC V5Z 1L3, Canada, [8]Sam & Ann Barshop Institute for Longevity & Aging Studies, UT Health San Antonio, San Antonio, TX 78229, USA, [9]Glenn Biggs Institute for Alzheimer's and Neurodegenerative Diseases, UT Health San Antonio, San Antonio, TX 78229, USA, [10]Faculty of Applied Science, University of British Columbia, Vancouver, BC V6T 1Z2, Canada, [11]Department of Molecular and Cellular Biology, UC Davis, Davis, CA 95616, USA and [12]May's Cancer Center, UT Health San Antonio, San Antonio, TX 78229, USA

Received June 29, 2022; Editorial Decision August 05, 2022; Accepted August 17, 2022

ABSTRACT

R-loops are three-stranded nucleic acid structures formed from the hybridization of RNA and DNA. In 2012, *Ginno et al.* introduced the first R-loop mapping method. Since that time, dozens of R-loop mapping studies have been conducted, yielding hundreds of publicly available datasets. Current R-loop databases provide only limited access to these data. Moreover, no web tools for analyzing user-supplied R-loop datasets have yet been described. In our recent work, we reprocessed 810 R-loop mapping samples, building the largest R-loop data resource to date. We also defined R-loop consensus regions and developed a framework for R-loop data analysis. Now, we introduce *RLBase*, a user-friendly database that provides the capability to (i) explore hundreds of public R-loop mapping datasets, (ii) explore R-loop consensus regions, (iii) analyze user-supplied data and (iv) download standardized and reprocessed datasets. *RLBase* is directly accessible via the following URL: https://gccri.bishop-lab.uthscsa.edu/shiny/rlbase/.

INTRODUCTION

R-loops are three stranded nucleic acid structures comprising an RNA:DNA hybrid and displaced single-stranded DNA (1). They occur as a byproduct of transcription at some genes and can be promoted by a variety of factors, including high G/C skew (2) and negative DNA superhelicity (3). Previous studies have implicated R-loops in a variety of pathological consequences, including their potential to promote replication stress (4), a phenomena observed in hypertranscriptional cancers such as Ewing sarcoma (5–7). Recent evidence has indicated that R-loops may also play important physiological roles in regulating gene expression (1), chromatin conformation (8) and ribosome biogenesis (9). However, the dynamics of R-loops, the mechanisms of their regulation, and the delineation between pathological and physiological R-loops remain largely unclear (1,10).

R-loop mapping studies apply high-throughput sequencing (HTS) to examine the causes and consequences of R-loops under differing biological conditions. In 2012, Ginno *et al.* introduced the first R-loop mapping technique, DNA:RNA immunoprecipitation sequencing (DRIP-Sequencing) (11). Since that time, >40 R-loop mapping studies have been conducted, yielding hundreds of publicly available R-loop samples (12). In our recent work, we mined 810 public R-loop mapping datasets, building the largest standardized R-loop data resource to date (12).

*To whom correspondence should be addressed. Tel: +1 210 562 9060; Email: bishopa@uthscsa.edu
Correspondence may also be addressed to Henry E. Miller. Tel: +1 210 562 9060; Email: millerh1@livemail.uthscsa.edu
Present address: Alexander Bishop, Greehey Children's Cancer Research Institute, Department of Cell Systems and Anatomy, UT Health San Antonio, San Antonio, TX 78229, USA.

From this work, we defined consensus sites of R-loop formation and developed a framework for the analysis of R-loop data. With these analysis and data resources, we created *RLBase*, a user-friendly database that provides the capability to (i) explore hundreds of R-loop mapping datasets, (ii) explore R-loop regions, (iii) analyze R-loop mapping datasets in the browser, and (iv) download processed and standardized R-loop mapping data.

Of note, other R-loop databases have been previously described (13–15). However, they provide only limited capability to access, explore and analyze R-loop data. The first, *R-loop DB*, enables the exploration of computationally predicted R-loop forming sequences (RLFS) along with 12 public R-loop mapping samples (14). However, it offers no capability to analyze user-supplied R-loop data, and data exploration is limited to a genome browser session. The second application, *R-loopAtlas*, is limited to plant genomes and, like *R-loop DB*, it only enables exploration via a genome browser session (13). The most recent web tool, *R-loopBase*, also only facilitates exploration of R-loop samples through a genome browser interface; it does not enable users to download the reprocessed R-loop datasets; it contains or analyze their own data; and while it is larger than previous databases, it does not include most R-loop mapping samples (15). Unlike all three previous web tools, *RLBase* provides full access to hundreds of reprocessed datasets, comprehensive tools for data exploration and in-browser data analysis capabilities.

Taken together, *RLBase* is a novel web database that provides R-loop biologists the valuable capability to access, explore and analyze R-loop mapping data in their web browser.

MATERIALS AND METHODS

RLBase data (upstream)

The processed data stored in *RLBase* was generated using a long-running computational pipeline available in its entirety in the *RLBase-data* GitHub repository (see *Code Availability*). The steps for upstream data processing were described in detail in our recent work (12). In addition to those details previously described, the present work will also describe the process of analyzing RNA-Sequencing datasets.

RNA sequencing datasets. In cases where R-loop mapping studies also provided matched RNA sequencing (RNA-Seq) datasets, they were downloaded and quantified to assess gene expression with *RLPipes* via the following procedure: raw reads were downloaded and pre-processed using the same procedure described in our recent work (12) up until the genomic alignment step and then reads were quantified using the *salmon* pseudo aligner (16) to generate read counts.

RLBase data (downstream)

Following generation of processed data files (peaks, coverage, expression quantification and quality statistics), downstream data processing was initiated to generate the final *RLBase* data. This involved (i) RLFS analysis, (ii) quality model building, (iii) sample classification, (iv) R-loop consensus analysis, (v) peak annotation and enrichment testing, (vi) Expression matrix generation, (vii) R-loop region annotation, (viii) sample-level correlation analysis, (ix) R-loop region abundance matrix generation, (x) calculating R-loop/expression correlation, (xi) updating the *RLBase* genome browser trackhub, (xii) RLSeq analysis of every sample and (xiii) upload of all data to the RLBase Amazon Web Services (AWS) S3 bucket. Notably, many of these steps (i-v,viii) are described in detail in our previous work (12). Therefore, the present work will describe only those (vi,vii,ix-xiii), which have not been described elsewhere.

Gene expression compilation. As part of the RLBase processing pipeline, RNA-Seq data from R-loop mapping studies were also analyzed. These data were quantified and saved to the disk (see *RLBase data (upstream)*). Then, the *buildExpression.R* script from the *RLBase-data* GitHub repository (see *Code Availability*) was executed to convert these quantification files to a *SummarizedExperiment* object, from the *SummarizedExperiment* R package (17). The procedure implemented in *buildExpression.R* followed these steps: (i) all quantification files were imported using the *tximport* function from the *tximport* R package (18) and summarized to the gene level. (ii) The transcripts per million (provided in each file) was log2 transformed. (iii) The variance-stabilizing transform (VST) of the data was calculated via the *vst* function from the *DESeq2* R package (19). The count matrix, log2TPM matrix and VST matrix were then combined using the *SummarizedExperiment* function (17) and saved to disk.

R-loop region annotation. R-loop regions (RL regions) were annotated via the *rlregionsToFeatures.R* script in the *RLBase-data* repository (see *Code availability*). For catalytically dead RNase H1 (dRNH), S9.6 and combined RL regions, the peaks were annotated with the hg38 genomic annotations previously described (see *Genomic features*) using the *bed_intersect* function from the *valr* R package (20).

R-loop region abundance calculation. R-loop region (RL region) abundance was calculated within each human sample in *RLBase* using the *rlregionCountMat.R* script from the *RLBase-data* GitHub repository (see *Code availability*). *RLBase* sample alignment files ('BAM' format) were processed with *featureCounts* from the *Rsubread* R package (21) to quantify the read counts from each within RL regions. Then log2 RL regions per million (log2RLRPM) was calculated using the same procedure for log2 transcript per million (log2TPM) calculation during gene expression analysis. Finally, the variance stabilizing transform (VST) was used to calculate normalized counts via the *vst* function from *DESeq2* (19). The three resulting matrices, raw counts, log2RLRPM and VST, were combined into a *SummarizedExperiment* object using the *SummarizedExperiment* R package (17) and saved to disk.

R-loop region abundance correlation with gene expression. Next, the *rlExpCorr.R* R script from the *RLBase-data* GitHub repository (see *Code availability*) was used to calculate the correlation of R-loop region (RL region) abundance. The procedure used is describe below.

Matching R-loop samples and expression samples. Some R-loop mapping studies also had matched RNA-Seq data. For each R-loop mapping sample, the 'study', 'tissue', 'genotype' and 'other' columns from the curated metadata were compared to the same column in the list of expression samples. If the values in all four columns were a match with at least one expression sample, then those four columns would be assigned as the 'exp_matchCond' (expression match condition). If only three were available, then they would become the 'exp_matchCond'. To see the order in which columns were checked for possible matches, view the *buildExpression.R* script in the RLBase-data GitHub repo (see *Code availability*).

Summarizing results within match conditions. Once match conditions were found, the R-loop abundance and gene expression data were summarized within them. Briefly, the R-loop abundance (log2RLRPM) was averaged within any R-loop mapping samples sharing the same 'exp_matchCond' to create one summarized log2RLRPM value within each exp_matchCond – RL Region pairing. For the expression samples, genes were mapped to RL regions, and the log2TPM was summed within them to create an RL region expression matrix. Then, the log2TPM within RL regions was averaged within any expression samples sharing an 'exp_matchCond' to yield one summarized log2TPM value within each exp_matchCond – RL Region pairing. Then, the Spearman correlation was calculated between log2RLRPM and log2TPM across exp_matchCond within each RL region to yield a correlation estimate (*Rho*) and *P* value within each RL region. *P* value adjustment was performed using the Benjamini–Hochberg procedure.

Building the RLBase genome browser session. The UCSC genome browser (22) provides a user-friendly way for exploring genomic data in a web browser interface. A TrackHub is a collection of genomic data that can be visualized and easily shared (23). To develop the *RLBase* TrackHub, the *buildGenomeBrowserHub.R* script from the *RLBase-data* GitHub repository was executed (see *Code availability*). Briefly, this script is used to build all the HTML descriptions for each human coverage track in *RLBase* along with a 'oneFile' that ensures they will be accessed correctly in the Genome Browser session. The oneFile was also augmented to include the RL regions, along with regions of high G or C skew (see *Genome annotations*), and RFLS (see *R-loop forming sequences*).

Running RLSeq and data upload. Finally, for each sample in *RLBase* which yielded called peaks, the *RLSeq* function was executed (see *RLSeq*) and HTML reports were generated. These steps were executed as part of running the *runRLSeq.R* script from the *RLBase-data* GitHub repository (see *Code availability*). The resulting *RLRanges* objects and *RLSeq* HTML reports were uploaded to the *RLBase* AWS S3 bucket along with all other processed data sets.

RLBase

RLBase is written in R shiny (24) and uses the *RLBase* Amazon Web Services (AWS) S3 bucket as a back-end storage solution. Notably, *RLBase* also includes extensive documentation that provides screenshots of all features with verbose descriptions, along with terminology explained, and FAQs. The following sections describe the features of *RLBase* and their methods.

RLBase datasets. *RLBase* contains standardized and reprocessed R-loop mapping data from 693 public samples (*RLBase* v1.0). The data were found via manual curation and reprocessed from raw sequencing reads using the *RLPipes* command-line tool (see *Code availability*). The data processing workflow is fully documented in our previous work (12). The resulting data were then uploaded to AWS S3 and made available publicly. The data and access methods are fully described in the 'Downloads' section of the *RLBase* web server. Following sample reprocessing, the data were quality-controlled using the quality control approach developed in our previous work (12), which classifies samples as 'POS' (expected to map R-loops) or 'NEG' (not expected to map R-loops). These classifications are also provided in the *RLBase* datastore and the models are accessible via the *RLBase* 'Downloads' page and through the *RLHub* R package (see *Code availability*). After quality control, high-confidence samples were analyzed to derive R-loop regions (RL regions), consensus sites of R-loop formation across the human genome. For details on the procedure used to produce these regions, see the description in our previous work (12). The RL regions can be downloaded via the *RLBase* 'Downloads' page or through the *RLHub* R package (see *Code availability*).

Samples. The 'Samples' page provides an interactive interface for exploring the 693 samples (RLBase v1.0) contained within *RLBase*. The interface for this tab is divided into three sections: (i) RLBase Samples Table, (ii) Table Controls and (iii) Outputs. The 'RLBase Samples Table' is an interactive, searchable, sortable, paginated table built with the *datatables* JavaScript library (via the *DT* R package) (25). It contains metadata for every sample in RLBase. Selecting a row in the table will update the 'Outputs' automatically. The 'Table Controls' are interactive user-interface (UI) elements, which control the data displayed in the table and in the outputs. The coordination between UI elements, table row selection and changes in the 'Outputs' is due to the built-in reactivity of R *shiny* (24). Caching is also employed to improve the performance of this interface. Finally, the 'Outputs' display a wealth of information, data, and plots that describe the samples (see *Results*, *RLBase*). The plots are produced with the built-in plotting functions in *RLSeq* (see *RLSeq plotting functions*) and with the *plotly* R package (26).

R-Loop regions. The 'R-Loop Regions' page provides an interactive interface for exploring the 64 418 R-loop regions (RL regions) uncovered in *RLBase* v1.0 (see *R-loop regions*). Like the 'Samples' page, it includes three sections which interact via the *reactivity* within *shiny* (24): (i) RL Regions Table, (ii) Table Controls and (iii) Outputs. The 'RL Regions Table' displays the RL regions and their metadata (see *R-loop Regions*). User-interface (UI) elements in the 'Table Controls' filter the RL Regions Table based on criteria

such as 'Repetitive' (overlaps with repetitive elements). The 'Outputs' show both an interactive summary of the RL region selected in the table and a plot showing the relationship between RL region abundance (log2RLRPM) and RL region expression (log2TPM) (see *R-loop regions*). Finally, this page includes links to an interactive UCSC genome browser session, which contains an interface for accessing all available coverage tracks and RL region signal tracks.

Analyze. The 'Analyze' page provides an in-browser interface for analyzing R-loop mapping data. The sample entry form is used to describe the sample metadata and upload peaks. Notably, no capability to upload coverage tracks is provided due to the size limitations of the server. Upon clicking the 'Start' button, a background R process is launched using the *r_bg* function from the *callr* R package (27). Then, the following steps are performed: (i) A UUID (universally unique identifier, via the *uuid* R package (28)) is assigned. (ii) The *knitr* R package (29) is used to render a template *Rhtml*, which produces an HTML progress page that is uploaded to the AWS S3 location for the user's sample and which the user is prompted to open. The progress indicator polls for updates using a custom *JavaScript* function and then it updates progressively as the analysis proceeds. (iii) A sweetalert (via the *shinyWidgets* R package (30)) is displayed to alert the user that their sample is processing and to show them the link for viewing the progress page. (iv) The *RLRanges* function from the *RLSeq* package is used to build the *RLRanges* object. (v) The *RLSeq* function is run to process all results for the sample (see *RLSeq*). (vi) The *report* function is implemented to knit an HTML report (see *report*). (vii) The finished *RLRanges*, *RLSeq* HTML report, and log files are all uploaded to the user's directory (identified only by UUID) in the AWS S3 bucket. (8) The user is alerted via sweetalert that their sample is ready. If there was an error, *knitr* is used to knit the *Rhtml* error template and this is uploaded instead of the results page along with all log files to aid in debugging. For more information on the usage and results of this analysis, see the *Results-RLBase* section.

Download. The *RLBase* 'Download' page provides convenient access to all processed data generated as part of the *RLBase-data* workflow (see *Code availability*). Instructions for bulk access via the AWS CLI are provided along with fine-grained access options for 'Processed data files', 'RLHub downloads' and 'Raw and misc data'. The 'Processed data files' tab provides an interactive, searchable table constructed with the *DT* R package (25) that provides download links for processed data files, such as peaks and coverage. The 'RLHub downloads' panel provides a table constructed with the *kableExtra* R package (31), which includes direct links for downloading the data objects in *RLHub* as well as the functions that can be used to access them via the *RLHub* R package. The 'Raw and misc data' panel provides an HTML-based guide to accessing raw and miscellaneous datasets.

Documentation. The documentation for *RLBase* was written in RMarkdown and knitted to an HTML file. The HTML is included in the *RLBase* web application via an *iframe*.

RESULTS

In our previous work, we reprocessed and standardized 810 R-loop mapping samples via a purpose-built computational pipeline, *RLPipes* (Figure 1A) (12). From meta-analysis of high-confidence R-loop samples, we also derived R-loop regions (RL regions), sites of consensus R-loop formation. We now present *RLBase,* a user-friendly webserver that facilitates the exploration of public R-loop mapping samples and RL regions, access to standardized and reprocessed datasets and the ability to analyze user-supplied R-loop mapping data (Figure 1B). In the following sections, we will describe the *RLBase* interface and core features.

Samples

The 'Samples' page (Supplementary Figure S1) provides the capability to explore the 810 publicly available R-loop mapping datasets provided by *RLBase*. These public R-loop mapping samples were reprocessed, standardized, quality controlled, analyzed with genomic feature enrichment analysis, and then used to derive R-loop regions (RL regions), sites of consensus R-loop formation (12). These analyses yielded a wealth of data that *RLBase* provides for exploration via the 'Samples' page (Supplementary Figure S1). Notably, this exploration is interactive as it contains user controls (Supplementary Figure S1A), an interactive data table (Supplementary Figure S1B) and responsive outputs (Supplementary Figure S1C) that provide the user with the capability to select samples and browse the visualizations and data relevant to them (Figure 2). The sample page output panel (Supplementary Figure S1C) contains: (i) The 'Summary' panel (Supplementary Figure S2), (ii) The 'Sample-sample comparison' panel (Supplementary Figure S3), (iii) The 'Annotation' panel (Supplementary Figure S4), (iv) The 'RLFS' panel (Supplementary Figure S5), (v) the 'RL Regions' panel (Supplementary Figure S6) and (vi) the sample 'Downloads' panel. The following will describe these outputs.

Summary. The summary panel provides a high-level overview of all samples in the 'RLBase Samples Table' (Supplementary Figure S2B-D) and of the specific sample selected in that table (Supplementary Figure S2A). It is used for examining the representation of various R-loop mapping modes, labels and predictions (Figure 2A) among the samples selected via the 'Table Controls' (Supplementary Figure S1A). From this interface, the user can also see the summary results for the sample which they select in the 'RLBase samples table' (Supplementary Figures S1A and S2A). From these results, the user can easily explore the data at a high level and perform univariate analyses.

Sample–sample comparison. The next panel, 'Sample–sample comparison' (Supplementary Figure S3), provides the user with visualizations that reflect the similarities and differences between the samples selected in the 'Table Controls' (Supplementary Figure S1A) and reveals the relationship between these samples and the single sample that is selected in the 'RLBase Samples Table' (Supplementary Figure S1B). The primary visualizations are a sample-level heatmap (Figure 2B and Supplementary Figure S3A) and

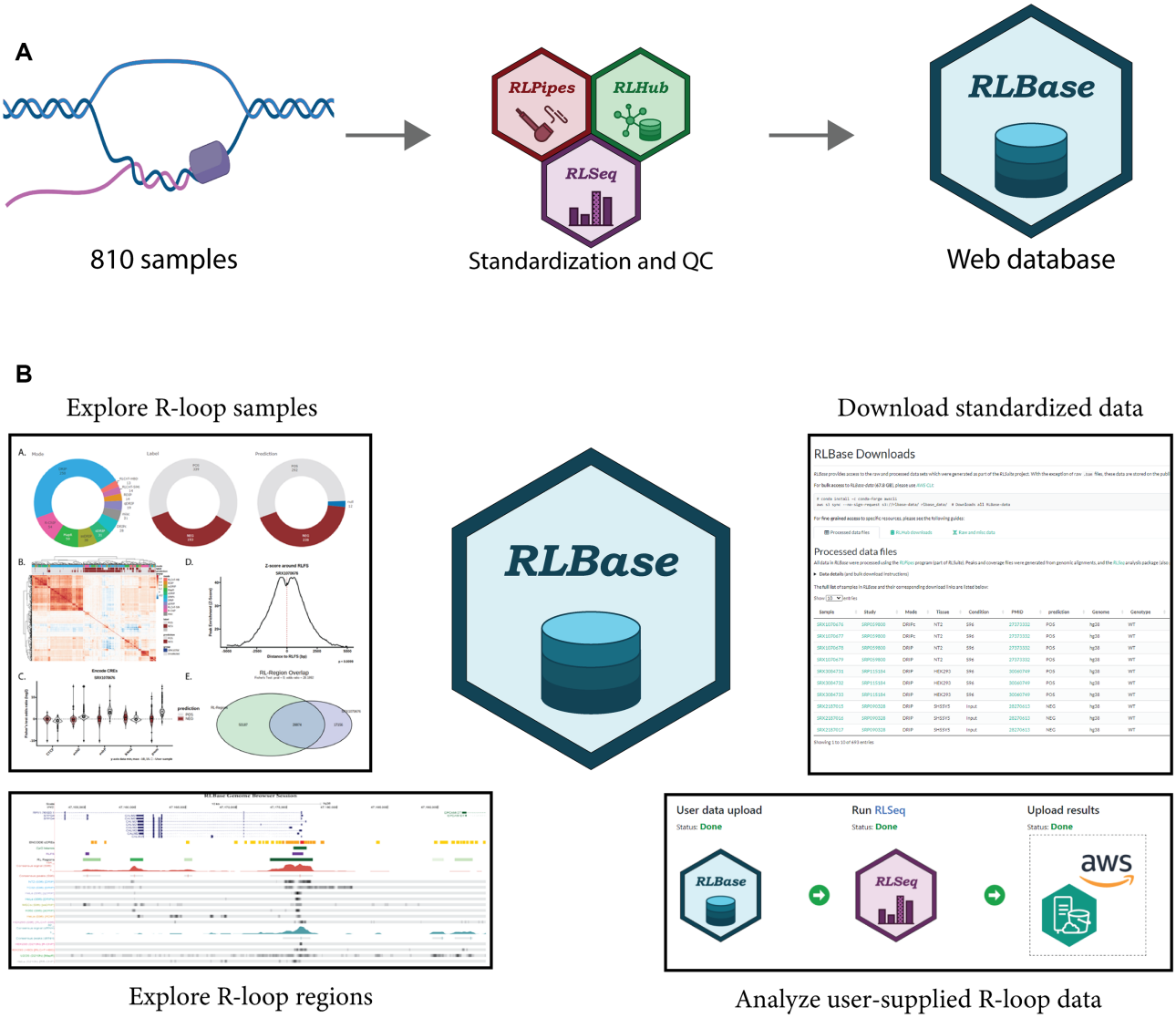

Figure 1. RLBase overview. (**A**) Graphical illustration showing the processing steps for RLBase. 810 R-loop datasets were downloaded and reprocessed using the *RLPipes*, *RLHub* and *RLSeq* software. *RLBase* is an R-shiny web database that was developed based on these software and data. (**B**) Graphical illustration depicting the core functionality of RLBase. RLBase provides the capability to (i) explore R-loop mapping samples and generate summary visualizations, (ii) explore R-loop regions and view them in the genome browser, (iii) analyze user-supplied R-loop mapping data in the browser and (iv) download standardized and reprocessed R-loop mapping data.

a PCA plot (Supplementary Figure S3B). These plots use the Pearson correlation between R-loop mapping samples around high-confidence R-loop sites to reveal which samples are most similar with each other and which are most different (12). These visualizations can be used to examine how certain modalities differ from one another and showcase the 'false positive' samples included in public R-loop mapping data (samples that are expected to map R-loops but likely do not). Moreover, they allow users to explore how an individual sample of interest relates to all other samples within the selected data. This can be useful for evaluating how a particular sample relates to those it should be most alike.

Annotation. The 'Annotation' panel (Supplementary Figure S4) showcases the results obtained from running the *RLSeq featureEnrich* function for each sample in the dataset (*see Availability*). The tabs in the output interface (Supplementary Figure S4B) control which annotations are displayed at any given time, the 'Table Controls' (Supplementary Figure S1A) control the background data present in the plot, and the selected row in the 'RLBase Samples Table' (Supplementary Figure S1B) controls which sample is highlighted in the plots (Supplementary Figure S4C and Figure 2C). Finally, the 'Split' selector controls whether the data are split by sample prediction, sample label or whether there is no split (Supplementary Figure S4A). The plots display the distribution of feature enrichment results across all samples selected, expressed in terms of the log2 odds ratio from Fisher's exact test (Figure 2C). These plots allow the user to assess both the differences found within the dataset across annotations and a spe-

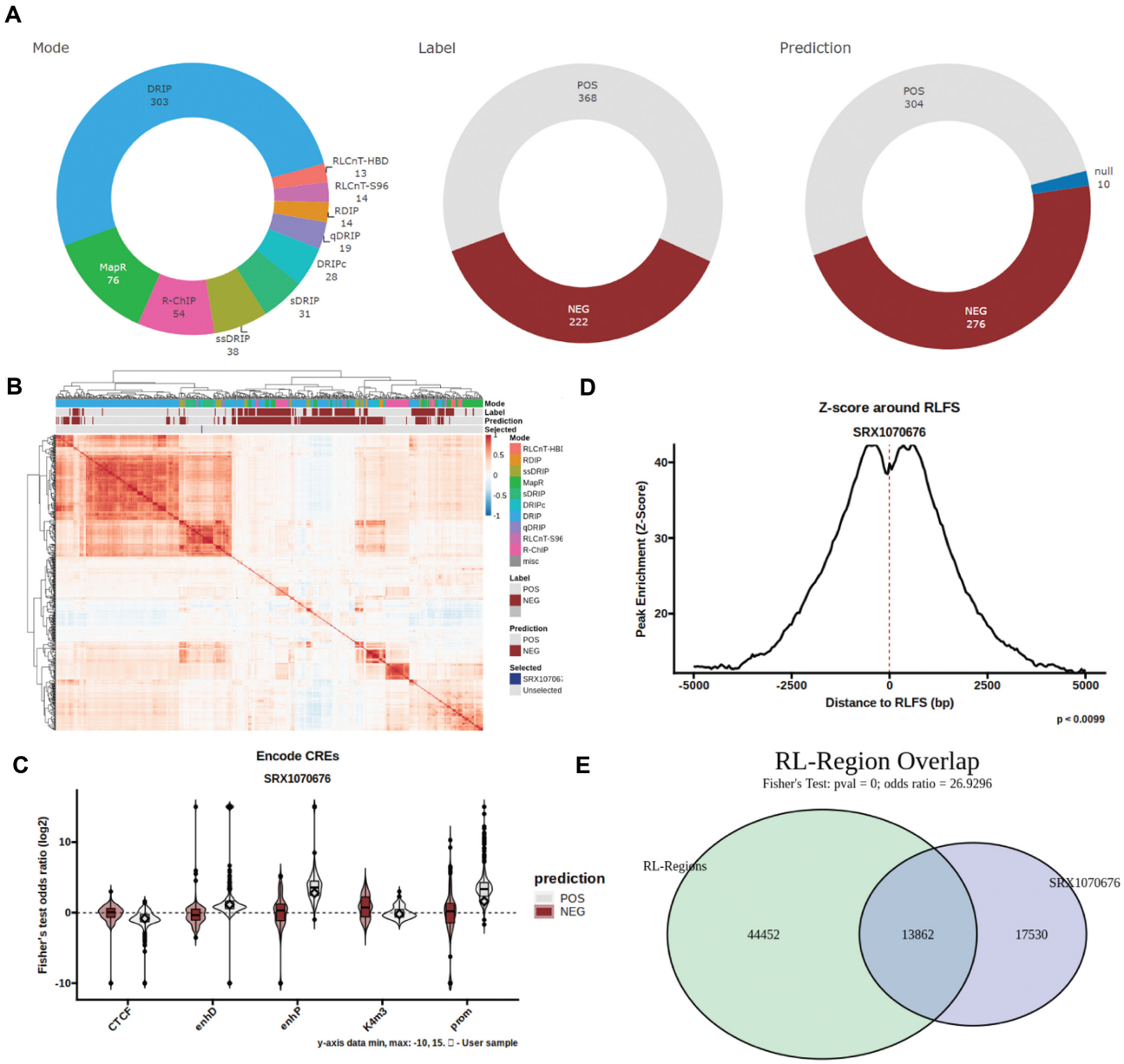

Figure 2. Select visualizations generated from RLBase. (**A–E**) Visualizations produced by RLBase with all modes, 'Show labeled controls' and 'Show predicted control' and 'hg38' genome options selected in the 'Table controls'. Additionally, the DRIPc sample SRX1070676 was selected in the 'RLBase Samples Table'. (A) Donut charts showing the proportions of samples by Mode, Label and Prediction. (B) The sample-sample correlation heatmap from the 'Sample-sample comparison' panel. (C) Genomic feature plots showing the distribution of ENCODE *cis*-regulatory element (CRE) feature enrichment within 'POS' (predicted to map R-loops) and 'NEG' (not predicted to map R-loops) samples. The user-selected sample (SRX1070676) is highlighted in each feature. Abbreviations: CTCF, CTCF binding site; enhD, distal enhancer; enhP, enhancer–promoter; K4m3, H3K4me3 histone modification site; prom, promoter site. (D) RLFS analysis plot with *P* value from permutation testing. (E) Venn diagram showing the overlap between selected sample ranges (SRX1070676 peaks) and RL regions. *P* value and odds ratio from Fisher's exact test.

cific sample of interest which they want to observe annotations for. These results reveal novel relationships between R-loops and genomic features and provide a useful quality metric (12).

RLFS. R-loop forming sequences (RLFS) are genomic regions favorable to the formation of R-loops (14,32). Moreover, they can be used to assess R-loop mapping sample quality (12). RLFS analysis was performed for each sample in *RLBase* via the *RLSeq analyzeRLFS* function. The resulting analysis plot, generated via the *RLSeq plotRLFSRes* function are provided for each sample in *RLBase* (Figure 2D and Supplementary Figure S5C). The 'RLFS' panel provides this plot along with other outputs that can be used to assess sample quality (Supplementary Figure S5): (i) a permutation testing plot (Supplementary Figure S5B) as obtained from the *plot* method within the *regioneR* package, (ii) an RLFS plot with Fourier trans-

form applied (Supplementary Figure S5D), which pertains to the features used by the quality model to render a quality prediction and (iii) a summary of the results from quality analysis (Supplementary Figure S5A). These results together give the user the ability to assess the quality of each sample.

RL regions. R-loop regions (RL regions) are consensus sites of R-loop formation discovered from the meta-analysis of high-confidence R-loop mapping samples (12). The overlap of each sample in *RLBase* with these RL regions was calculated via the *RLSeq rlRegionTest* function. The visualization and full data table associated with these results is provided in the 'RL Regions' panel (Supplementary Figure S6). The visualization is a Venn diagram showing the overlap of the peaks in the user-selected sample with the RL regions (Figure 2E and Supplementary Figure S6A). The *P* value and odds ratio are derived from the Fisher's exact test. The RL region table is also provided (Supplementary Figure S6B), which lists all the RL regions uncovered by the peaks in the user-selected sample. These results provide the user with the capability to explore the degree of overlap between RL regions with the sample they select. It also allows the user to view the specific RL regions, which were uncovered in the selected sample.

R-loop regions

The 'R-Loop Regions' tab (Supplementary Figure S7) provides the tools necessary to explore consensus sites of R-loop formation (RL regions) derived from high-confidence *RLBase* samples (12). These regions contain a wealth of metadata, including their ID, genomic location, associated genes and confidence level (Supplementary Figure S7B). The user is provided with 'Table Controls' (Supplementary Figure S7A) that allow them to control which RL regions are displayed and how gene names are shown. The 'All genes' checkbox allows the user to control whether all genes overlapping an RL region are displayed, or whether only well-annotated genes are shown (those which appear in pathway databases). The 'Repetitive regions' checkbox controls whether to show RL regions which overlap with repetitive genomic sequences, such as pericentromeric regions. Finally, the 'Correlated with expression' checkbox will, if selected, only show the RL regions for which a significant correlation between R-loop abundance and gene expression was observed. The output panel (Supplementary Figure S7D) displays a summary of the RL region currently selected in the 'RL Regions Table' (Supplementary Figure S7B) and displays a correlation plot showing the relationship between R-loop abundance and gene expression levels.

Finally, the *RLBase* UCSC genome browser session is provided (Figure 3 and Supplementary Figure S7C). This browser session provides access to all the human R-loop mapping samples in *RLBase* along with consensus signal, peaks, RL regions, RLFS and other relevant annotations. Taken together, these features provide the user with the capability to explore RL regions, observe their association with expression and view them alongside all other datasets in *RLBase* in the UCSC genome browser.

Analyze

The 'Analyze' page (Supplementary Figure S8) provides users with the ability to analyze their own R-loop mapping data in the browser without the need for bioinformatics skill. To analyze their data, a user provides metadata about the sample to be analyzed (Supplementary Figure S8A), the peaks (in BED format) for the sample (Supplementary Figure S8B) and agrees to the privacy policy (Supplementary Figure S8C). Having performed these steps, the user selects 'Start' and launches the analysis job. The user is presented with the results link, which will show the progress of the analysis and, when completed, will update to show the analysis results (Supplementary Figure S9). The results page contains a verbose description of the analysis (Supplementary Figure S9A), the progress indicator (Supplementary Figure S9B), and a table with sample metadata and the results (once available) (Supplementary Figure S9C). These results contain the report which details the analysis results, the *RLRanges* R data object which the user can load in an R session for further analysis, and the logs generated by the analysis, which are useful for debugging should an error be encountered. Taken together, the 'Analyze' page provides users with an easy and convenient in-browser interface to the functions used for analysis, yielding quality reports which have a permalink suitable for sharing.

Download and documentation

The 'Download' page provides direct links to download all processed files associated with *RLBase*, including verbose descriptions of each item and of programmatic methods for accessing these data. It also includes instructions for how to download the alignment (BAM) files for each sample and how to access processed datasets via the *RLHub* R/Bioconductor package. Finally, the 'Documentation' page provides verbose descriptions of features and addresses terminology and technical concepts, which may need additional explanation.

LIMITATIONS AND FUTURE DIRECTIONS

RLBase provides a user-friendly web interface for the exploration, access and analysis of R-loop datasets. However, there are several key limitations which exist in the current version of the database. *Tissue types, modality and species of origin:* One key limitation of *RLBase* is that many modalities, tissue types and species contain few samples representing them. This leads to biases that may degrade the utility of the exploration and analysis tools for some use cases. As discussed in our recent work (12), we addressed this limitation by developing an online learning scheme that enables rapid incorporation of new R-loop data as they become available. With this scheme, we intend on updating RLBase regularly to incorporate these new data. Over time, this approach will improve the robustness of the RLBase analysis and exploration features for those use cases. *Analysis of R-loop changes under perturbation conditions:* The current database contains a variety of R-loop samples from both basal (e.g. untreated samples) and perturbed conditions (e.g. drug-treated samples). No distinction is currently made between these two types of samples. Moreover, no

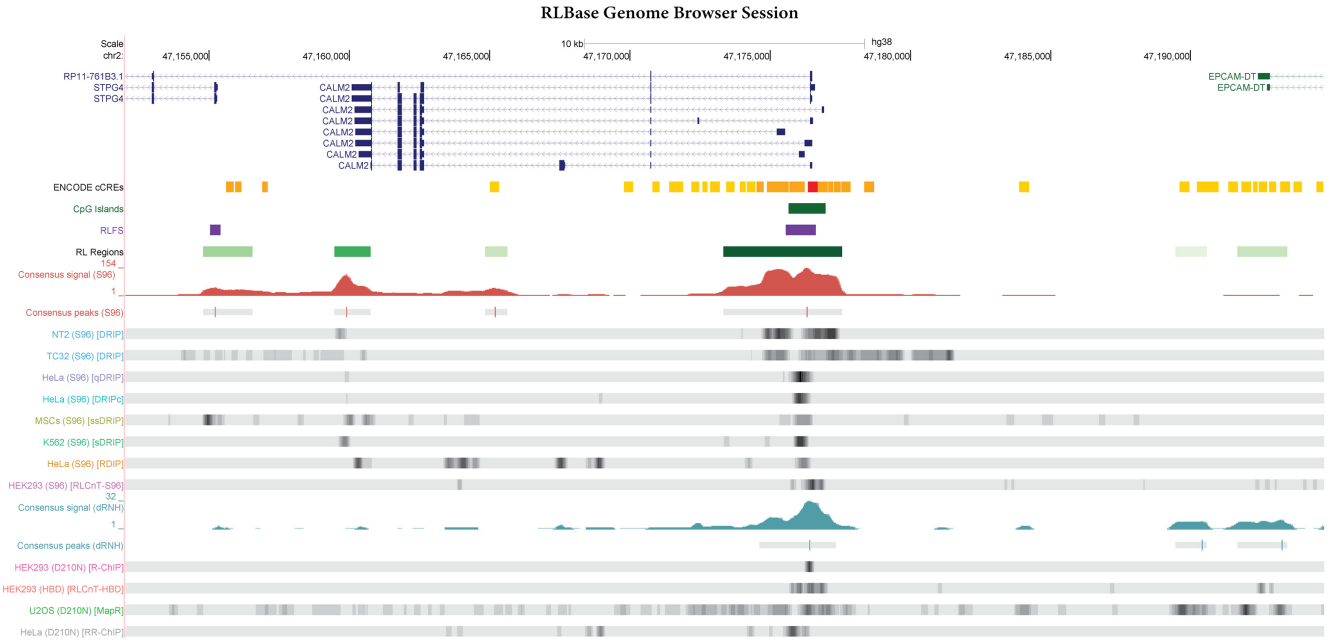

Figure 3. RLBase genome browser session. The browser session includes representative RLBase sample coverage tracks alongside S9.6 and dRNH consensus signal, RL regions, CpG islands, R-loop forming sequences (RLFS) and ENCODE CREs. The screen capture shown here was taken in the area surrounding the CALM2 gene.

tools are currently provided to enable the exploration of how these perturbations impact R-loop dynamics, despite the utility of such an analysis. It will remain for future versions of RLBase to provide these capabilities, especially as more data becomes available with larger numbers of diverse perturbations.

CONCLUSION

RLBase is a user-friendly web database for the exploration and analysis of R-loop data. It provides the capability to (i) explore reprocessed and standardized R-loop mapping datasets, (ii) explore R-loop regions (consensus sites of R-loop formation), (iii) analyze user-supplied R-loop datasets, and (iv) download re-processed and standardized R-loop data.

DATA AVAILABILITY

Data availability. All data are made available through the *RLHub* R/Bioconductor package and the *RLBase* web interface. *RLBase* URL: https://gccri.bishop-lab.uthscsa.edu/rlbase/. *RLHub* URL: https://bioconductor.org/packages/devel/data/experiment/html/RLHub.html.

Code availability. All software developed as part of this project are available publicly under an MIT license. Data processing scripts: https://github.com/Bishop-Laboratory/RLBase-data. RLBase source code: https://github.com/Bishop-Laboratory/RLBase. Additional software packages used include (i) RLSeq (available via Bioconductor): https://bioconductor.org/packages/devel/bioc/html/RLSeq.html, (ii) RLHub (available via Bioconductor): https://bioconductor.org/packages/devel/data/experiment/html/RLHub.html and (3) RLPipes (available via Bioconda): https://anaconda.org/bioconda/rlpipes.

All accessions for data used in *RLBase* are listed in Supplementary Table S1 of our previous publication (12).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We want to thank the Bishop Laboratory for their insightful comments and feedback on the web database interface. We want to thank Lori Kern at Bioconductor for her review of *RLSeq* and *RLHub*. We want to also thank Simon Bray at Bioconda for his review of *RLPipes*. We want to thank the Bioinformatics Research Network for their support in developing *RLBase*. Figure 1A was created, in part, using BioRender.com.

FUNDING

NIH/NCI [R01CA152063, 1R01CA241554 to A.J.R.B], CPRIT [RP150445 to A.J.R.B]; Research sponsored by a SU2C-CRUK Pediatric Cancer New Discoveries Challenge Team Grant (Grant Number: SU2C #RT6187 to A.J.R.B); Greehey Graduate Fellowship Award to H.E.M.; NIH/NIA [F31AG072902 to H.E.M.]; NIH [R35 GM139549 to F.C.]; DOD [CDMRP PR181598 to K.S.]. Funding for open access charge: National Institutes of Health.

Conflict of interest statement. None declared.

REFERENCES

1. Niehrs,C. and Luke,B. (2020) Regulatory R-loops as facilitators of gene expression and genome stability. *Nat. Rev. Mol. Cell Biol.*, **21**, 167–178.
2. Ginno,P.A., Lim,Y.W., Lott,P.L., Korf,I. and Chédin,F. (2013) GC skew at the 5′ and 3′ ends of human genes links R-loop formation to epigenetic regulation and transcription termination. *Genome Res.*, **23**, 1590–1600.
3. Chedin,F. and Benham,C.J. (2020) Emerging roles for R-loop structures in the management of topological stress. *J. Biol. Chem.*, **295**, 4684–4695.
4. Cristini,A., Groh,M., Kristiansen,M.S. and Gromak,N. (2018) RNA/DNA hybrid interactome identifies DXH9 as a molecular player in transcriptional termination and R-Loop-Associated DNA damage. *Cell Rep.*, **23**, 1891–1905.
5. Gorthi,A., Romero,J.C., Loranc,E., Cao,L., Lawrence,L.A., Goodale,E., Iniguez,A.B., Bernard,X., Masamsetti,V.P., Roston,S. *et al.* (2018) EWS-FLI1 increases transcription to cause R-loops and block BRCA1 repair in Ewing sarcoma. *Nature*, **555**, 387–391.
6. Miller,H.E., Gorthi,A., Bassani,N., Lawrence,L.A., Iskra,B.S. and Bishop,A.J.R. (2020) Reconstruction of Ewing sarcoma developmental context from mass-scale transcriptomics reveals characteristics of EWSR1-FLI1 permissibility. *Cancers (Basel)*, **12**, E948.
7. Gorthi,A. and Bishop,A.J.R. (2018) Ewing sarcoma fusion oncogene: at the crossroads of transcription and DNA damage response. *Mol. Cell Oncol.*, **5**, e1465014.
8. Pan,H., Jin,M., Ghadiyaram,A., Kaur,P., Miller,H.E., Ta,H.M., Liu,M., Fan,Y., Mahn,C., Gorthi,A. *et al.* (2020) Cohesin SA1 and SA2 are RNA binding proteins that localize to RNA containing regions on DNA. *Nucleic Acids Res.*, **48**, 5639–5655.
9. Abraham,K.J., Khosraviani,N., Chan,J.N.Y., Gorthi,A., Samman,A., Zhao,D.Y., Wang,M., Bokros,M., Vidya,E., Ostrowski,L.A. *et al.* (2020) Nucleolar RNA polymerase II drives ribosome biogenesis. *Nature*, **585**, 298–302.
10. Castillo-Guzman,D. and Chédin,F. (2021) Defining R-loop classes and their contributions to genome instability. *DNA Repair (Amst)*, **106**, 103182.
11. Ginno,P.A., Lott,P.L., Christensen,H.C., Korf,I. and Chédin,F. (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol. Cell*, **45**, 814–825.
12. Miller,H.E., Montemayor,D., Abdul,J., Vines,A., Levy,S.A., Hartono,S.R., Sharma,K., Frost,B., Chédin,F. and Bishop,A.J.R. (2022) Quality-controlled R-loop meta-analysis reveals the characteristics of R-loop consensus regions. *Nucleic Acids Res.*, **50**, 7260–7286.
13. Xu,W., Li,K., Li,S., Hou,Q., Zhang,Y., Liu,K. and Sun,Q. (2020) The R-Loop atlas of arabidopsis development and responses to environmental stimuli. *Plant Cell*, **32**, 888–903.
14. Jenjaroenpun,P., Wongsurawat,T., Sutheeworapong,S. and Kuznetsov,V.A. (2017) R-loopDB: a database for R-loop forming sequences (RLFS) and R-loops. *Nucleic Acids Res.*, **45**, D119–D127.
15. Lin,R., Zhong,X., Zhou,Y., Geng,H., Hu,Q., Huang,Z., Hu,J., Fu,X.-D., Chen,L. and Chen,J.-Y. (2022) R-loopBase: a knowledgebase for genome-wide R-loop formation and regulation. *Nucleic Acids Res.*, **50**, D303–D315.
16. Patro,R., Duggal,G., Love,M.I., Irizarry,R.A. and Kingsford,C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
17. Morgan,M., Obenchain,V., Hester,J. and Pagès,H. (2021) SummarizedExperiment: summarizedexperiment container. https://doi.org/10.18129/B9.bioc.SummarizedExperiment.
18. Love,M., Soneson,C., Robinson,M., Patro,R., Morgan,A.P., Thompson,R.C., Shirley,M. and Srivastava,A. (2021) tximport: import and summarize transcript-level estimates for transcript- and gene-level analysis. https://doi.org/10.18129/B9.bioc.tximport.
19. Love,M., Ahlmann-Eltze,C., Forbes,K., Anders,S. and Huber,W., (2021) DESeq2: differential gene expression analysis based on the negative binomial distribution. https://doi.org/10.18129/B9.bioc.DESeq2.
20. Riemondy,K.A., Sheridan,R.M., Gillen,A., Yu,Y., Bennett,C.G. and Hesselberth,J.R. (2017) valr: reproducible genome interval analysis in R. *F1000Res*, **6**, 1025.
21. Liao,Y., Smyth,G.K. and Shi,W. (2019) The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res.*, **47**, e47.
22. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
23. Raney,B.J., Dreszer,T.R., Barber,G.P., Clawson,H., Fujita,P.A., Wang,T., Nguyen,N., Paten,B., Zweig,A.S., Karolchik,D. *et al.* (2014) Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics*, **30**, 1003–1005.
24. Chang,W., Cheng,J., Allaire,J.J., Sievert,C., Schloerke,B., Xie,Y., Allen,J., McPherson,J., Dipert,A., Borges,B. *et al.* (2021) shiny: web application framework for R.
25. Xie,Y., Cheng,J., Tan,X., Allaire,J.J., Girlich,M., Ellis,G.F. and Rauh,J.(2021) DT: A Wrapper of the JavaScript Library 'DataTables'. https://rdrr.io/cran/DT/.
26. Sievert,C., Parmer,C., Hocking,T., Chamberlain,S., Ram,K., Corvellec,M., Despouy,P., Brüggemann,S. and Inc,P.T. (2021) plotly: create interactive web graphics via 'plotly.js'.
27. Csárdi,G. and Chang,W. (2021) callr: Call R from R. https://cran.r-project.org/web/packages/callr/index.html.
28. Urbanek,S. (2020) uuid: Tools for Generating and Handling of UUIDs. https://rdrr.io/cran/uuid/man/UUIDgenerate.html.
29. Xie,Y., Sarma,A., Vogt,A., Andrew,A., Zvoleff,A., Atkins,A., Wolen,A. and Manton,A. (2021) knitr: A General-Purpose Package for Dynamic Report Generation in R. https://rdrr.io/github/yihui/knitr/man/knitr-package.html.
30. Perrier,V., Meyer,F. and Granjon,D. (2021) shinyWidgets: Custom Inputs Widgets for Shiny. https://rdrr.io/cran/shinyWidgets/.
31. Zhu,H., Travison,T., Tsai,T., Beasley,W., Xie,Y., Yu,G., Laurent,S., Shepherd,R., Sidi,Y. *et al.* (2021) kableExtra: construct complex table with 'kable' and pipe syntax. https://rdrr.io/cran/kableExtra/.
32. Jenjaroenpun,P., Wongsurawat,T., Yenamandra,S.P. and Kuznetsov,V.A. (2015) QmRLFS-finder: a model, web server and stand-alone tool for prediction and analysis of R-loop forming sequences. *Nucleic Acids Res.*, **43**, W527–W534.