Quality-controlled R-loop meta-analysis reveals the characteristics of R-Loop consensus regions

Miller, H. E.[1,2,3,*], Montemayor D.[4,5], Abdul, J.[3,5], Vines, A.[3,7], Levy, S.[1,3,8,9], Sharma, K.[4,5], Frost, B. [1,8,9], Chedin, F.[10], Bishop, A. J. R.[1,2]

[1]Department of Cell Systems and Anatomy, UT Health San Antonio, San Antonio, TX, USA

[2]Greehey Children's Cancer Research Institute, UT Health San Antonio, San Antonio, TX, USA

[3]Bioinformatics Research Network, Atlanta, GA, USA

[4]Department of Medicine, UT Health San Antonio, San Antonio, TX, USA

[5]Center for Renal Precision Medicine, UT Health San Antonio, San Antonio, TX, USA

[6]Department of Biology, University of Ottawa, Ottawa, Canada,

[7]Faculty of Arts, University of Bristol, Bristol, U.K.

[8]Sam & Ann Barshop Institute for Longevity & Aging Studies, UT Health San Antonio, San Antonio, TX, USA

[9]Glenn Biggs Institute for Alzheimer's and Neurodegenerative Diseases, UT Health San Antonio, San Antonio, TX, USA

[10]Department of Molecular and Cellular Biology, UC Davis, Davis, CA, USA

[*]Address correspondence to millerh1@uthscsa.edu

Abstract

R-loops are three-stranded nucleic acid structures formed from the hybridization of RNA and DNA during transcription. While the pathological consequences of R-loops have been well-studied to date, the locations, classes, and dynamics of physiological R-loops remain poorly understood. R-loop mapping studies use high-throughput sequencing to reveal R-loop locations and dynamics genome-wide. While individual mapping studies provide insight into the role of R-loops under specific biological conditions, their findings are challenging to generalize. This is partially due to the narrow biological scope of individual studies, but it is also due to the limitations that affect each mapping modality and, in some cases, poor quality data which goes undetected. In this study, we reprocessed 693 R-loop mapping datasets from a wide array of biological conditions and mapping modalities. From this data resource, we developed a novel method for quality assessment, one which shows excellent accuracy and validity. In applying this method to our reprocessed and standardized data, we reveal the extent of poor-quality R-loop mapping datasets within previously published studies while also defining a high-confidence dataset suitable for meta-analysis. From these high-confidence samples, we proceed to define consensus R-loop sites called "R-loop regions" (RL regions). In the process, we reveal the stark divergence between S9.6 and dRNH-based R-loop mapping methods and identify biologically meaningful subtypes of both constitutive and variable R-loops. This work provides a much-needed method to assess R-loop data quality and we show that its implementation reveals intriguing aspects of R-loop biology.

Introduction

R-loops are three-stranded nucleic acid structures formed from the hybridization of RNA and DNA, often in regions with high G (guanine) or C (cytosine) skew, such as CpG islands (1–3). The study of R-loops has long focused on their pathological consequences in promoting genomic instability (4, 5) and in hypertranscriptional cancers, such as Ewing sarcoma (6, 7). Recent studies have also revealed that R-loops contribute to physiological processes, such as DNA repair (8), DNA methylation (9), and even reprogramming to pluripotency (10). While these studies have led to an increase in interest regarding the cause and consequences of

physiological R-loops, fundamental questions regarding their subtypes, dynamics, and interactions remain unanswered.

R-loop mapping via high-throughput sequencing can be utilized to elucidate R-loop locations and dynamics genome-wide (2). Presented in 2012, DNA:RNA immunoprecipitation (DRIP) sequencing was the first such technique, utilizing the S9.6 antibody to detect R-loops (2). Through DRIP sequencing, it was first demonstrated that R-loops form in many, but not all, genes (2). In recent years, improvements to DRIP-Seq have been established, such as DRIPc-Seq, which also uses the S9.6 antibody (11). In 2017, *Chen et al.* introduced "R-ChIP", a modality for R-loop mapping which relies on catalytically inactive RNase H1 (dRNH) as an alternative to S9.6 (12). Though preliminary evidence suggested differences between dRNH and S9.6-based R-loop mapping exist (12, 13), the extent and biological relevance of these disparities remains unclear. Recently, *Castillo-Guzman and Chedin* reviewed the relevant R-loop mapping literature and hypothesized that dRNH and S9.6 may map biologically distinct subtypes of R-loops ("Class I versus Class II") (14), but no studies to date have evaluated these subtypes bioinformatically. It is also unknown how different R-loop mapping modalities relate to each other more generally or whether they each reliably map the same sets of R-loops. This gap in knowledge has become a critical concern as, at the time of writing, there are now 23 distinct R-loop mapping modalities, 14 of which were developed since 2019 alone (**Table S1**).

In 2020, we reprocessed and standardized 108 public DRIP-Seq datasets to perform a meta-analysis of R-loop locations in relation to the cohesin complex (15). This initial mining study demonstrated the potential for greater insight from larger and more diverse datasets. However, it also revealed concerning inconsistencies in quality between publicly available R-loop mapping samples (15). These findings were echoed by the recent work of *Chedin et al. 2021*, who demonstrated that issues of data quality persist even within prominent R-loop mapping studies (16). While our work and that of *Chedin et al.* indicates the need for robust quality control methods in R-loop data analysis, no such techniques have yet been proposed to our knowledge.

In the present study, we profiled 693 publicly available R-loop mapping datasets. We devised methods for the quality control of these data and applied these methods to select high-quality R-loop mapping samples. Finally, we defined consensus sites of R-loop formation, termed "R-loop regions" (RL regions). We use these regions to elucidate the stark contrasts between S9.6 and dRNH-based R-loop mapping results, supporting the proposed "Class I/II" distinction and revealing differing biological programs among the R-loops uncovered by each mapping approach. Moreover, we define constitutive and variable R-loop regions, finding that they differ with respect to pathways relevant for transcription and replication.

In this study, we also demonstrate new approaches for the quality control and analysis of R-loop mapping data, which we provide in the collection of software tools, "RLSuite", which accompany this work (see *Code Availability*).

Methods

R-loop forming sequences (RLFS)

R-loop forming sequences were computationally predicted using the *QmRLFS-finder.py* python program (17) implemented as part of the *makeRLFSBeds.R* script in the accompanying data generation repository (see *Code Availability*). Briefly, every genome for which gene annotations were available was downloaded in "FASTA" format from the University of California Santa Cruz (UCSC) genome repository. Then the *QmRLFS-finder.py* script was run to generate predictions. The results table was then converted to "BED" format and saved.

Cataloging R-loop mapping datasets

R-loop data was found by querying the Gene Expression Omnibus (GEO), Sequence Read Archive (SRA), ArrayExpress, and PubMed with the keywords "R-loop", "R-loops", "RNA:DNA hybrid", "RNA:DNA hybrids", "RNA-DNA hybrid", "RNA-DNA hybrids", "DRIP-Seq", "R-ChIP", "S9.6", and "D210N". Query results with R-loop mapping data publicly accessible via SRA were included. For every R-loop mapping sample, the following details were recorded:

1. The sample accession (GEO or SRA).
2. The accession for the corresponding genomic input sample (if available).
3. The sample condition (e.g., RNase H1 treated, "RNH").
4. The mapping mode (e.g., "DRIP" for DNA-RNA immunoprecipitation sequencing (DRIP-Seq)).

5. The tissue of the sample. For cell lines, this was the cell line name (e.g., “HEK293”). For organs and whole organisms, the tissue name (e.g., “B-cell” for B-cells, “SC” for whole brewer’s yeast (*Saccharomyces cerevisiae*)).
6. The sample genotype, where relevant (e.g., “BRCA1” for samples derived from BRCA1 mutant breast cancer patients).
7. “Other”, a catchall for other relevant metadata provided by the original data submitters, such as drug treatment information.
8. The source publication PMID, if available.
9. The last author on the source publication, if available.

Please refer to **Table S1** for the full list of samples.

Standardization of sample metadata

The manually curated catalog was then standardized via the custom scripts in the accompanying data generation repository (see *Code Availability*). First, the “condition” was binarized to one of the following labels: “POS” (samples which were expected to map R-loops, e.g., “D210N” in R-ChIP data) or “NEG” (samples which were not expected to map R-loops, e.g., “Input” for DRIP-Seq data). The mapping of “condition” to “label” relied on a manually curated regular expressions (REGEX) dictionary and simple string matching (**Table S1**). Additionally, the manifest was filtered to keep only R-loop mapping samples (except for bisulfite sequencing samples). Finally, the *RLPipes* v0.9.0 software program was used to query the SRA database for missing metadata related to each sample (see *Code Availability*).

R-loop mapping data standardization

To reprocess all available R-loop mapping datasets, we used a long-running computational pipeline that is available in its entirety with detailed instructions in the accompanying data generation repository (see *Code Availability*). For all upstream processing steps, *RLPipes* v0.9.0 was used (see *Code Availability*). This pipeline was executed on a remote server running Ubuntu 18.04 with 500GB of RAM, 192 cores, and a 10GiB/s internet connection available. The pipeline took one week to finish processing all samples. The primary steps of the pipeline involved all the following:

First, raw reads in SRA format were downloaded for each SRA run via the *prefetch* software from NCBI *sra-tools*. Then, reads were converted to “FASTQ” format using *fastq-dump* from *sra-tools*. Next, technical replicates were merged and interleaved (in the case of paired-end data) using *reformat.sh* from *bbtools* (18). Then, reads were trimmed and filtered with *fastp* (19), generating a quality report.

For R-loop mapping data, reads were aligned to the appropriate genome using *bwa-mem2* (20). Then, alignments were filtered (minimum quality score of 10 enforced), sorted, and indexed using *samtools* (21) and duplicates were marked using *samblaster* (22). Then, peaks were called using *macs3* (23) and coverage was calculated using *deepTools* (24).

The outputs of the pipeline were (A) peaks in “broadPeak” format, (B) coverage in “bigWig” format, (C) quality reports for alignments and raw reads, and (D) alignments in “BAM” format (see *Data Availability*).

Normalization of read coverage

To calculate normalized coverage tracks for visualization purposes, we first calculated scaling factors using the trimmed mean of M-values (TMM) approach previously demonstrated for similar sequencing approaches (25) implemented via the *edgeR* R package (26). Then, we calculated the expected fragment sizes for each sample using the *macs3 predictd* method (23). We then calculated coverage using the *bamCoverage* command from *deepTools* (24) with arguments “--scaleFactor <scale_factor> -e <fragment_size> –ignoreDuplicates” where “<scale_factor>” is the scaling factor derived from *edgeR* and “<fragment_size>” is the predicted fragment size from *macs3*. The resulting tracks were visualized in the UCSC genome browser.

R-loop mapping data quality control

Following data standardization, a quality control model was developed to filter out poor-quality samples prior to meta-analysis of R-loop consensus sites. This involved (A) R-loop forming sequences (RLFS) analysis, (B) quality model building, (C) sample classification, and (D) model validation.

R-loop forming sequences analysis (A)

R-loop forming sequences (RLFS) analysis was performed using a custom R script, *rlfsAnalyze.R*, available in the accompanying data generation repository (see *Code Availability*). The analysis script calls the *analyzeRLFS*

function from the *RLSeq* R package (see *Code Availability*). This function implements a data-agnostic method for assessing the enrichment of R-loop mapping peaks within RLFS via the following procedure:

Permutation testing is implemented via the *permTest* function from the *regioneR* (27) R package such that, for each permutation, R-loop peaks were randomized using the *circularRandomizeRegions* function and then the number of overlaps with RLFS was counted. 100 permutations are used by default to build an empirical null distribution for peak/RLFS overlap. Then the true number of overlaps from non-randomized peaks and RLFS was compared to the null distribution to calculate the Z-score and significance of enrichment. Finally, a Z-score distribution was calculated using the *localZScore* function from *regioneR* (27) within 5kb upstream and downstream of the meta-RLFS midpoint. For more detail, see the *RLSeq* reference (see *Code availability*). For visualization purposes, the scaled Z score distributions were calculated for each peak set in the reprocessed data and visualized using locally estimated scatterplot smoothing (LOESS) regression with the standard error as the confidence interval.

Quality model building (B)

The quality classification model was built using the *models.R* script in the accompanying data generation repository (see *Code Accessibility*). The process of model building is semi-supervised and utilizes a graphical user interface to allow operators with minimal coding experience to conveniently use it. The model building procedure involves the following steps: (A) The operator executes the *models.R* script from the command line, launching an interactive web interface that contains a table with SRA experiment ID and label ("POS" or "NEG") for each sample along with a plot showing the Z-score distribution from the RLFS analysis of that sample. This interface also indicates samples that have been previously deliberated about. (B) The operator then decides which samples to exclude from the model building process. A sample should only be excluded when the Z-score distribution drastically differs from the expected Z-score distribution for its label (examples are provided to the operator in the web interface). (C) Once all samples with a label mismatch are selected, the operator clicks the "Build model" button which executes the model building process on the non-exculded data via the RMarkdown notebook, *FFT-classifier.Rmd* (available from the accompanying data generation repository). This notebook automatically executes the following steps: (C.1) The data are wrangled from the *RLRanges* objects. (C.2) Then, the following engineered features are calculated from the Z-score distributions (**Table 1**).

Feature	Description
Z1	mean of Z
Z2	variance of Z
Zacf1	mean of ACF(Z)
Zacf2	variance of ACF(Z)
ReW1	mean of FT(Z) (real part)
ReW2	variance of FT(Z) (real part)
ImW1	mean of FT(Z) (imaginary part)
ImW2	variance of FT(Z) (imaginary part)
ReWacf1	mean of FT(ACF(Z)) (real part)
ReWacf2	variance of FT(ACF(Z)) (real part)
ImWacf1	mean of FT(ACF(Z)) (imaginary part)
ImWacf2	variance of FT(ACF(Z)) (imaginary part)

Table 1. Engineered features used in the quality model. Where "Z" is Z-score distribution, "ACF" is the autocorrelation function, and "FT" is the Fourier Transform (calculated via the fft function in R).

The engineered features are then wrangled into matrix form with *samples* in rows and *engineered features* and *label* as the columns. The *label* in this case is "POS" (expected to map R-loops) or "NEG" (not expected to map R-loops).

(C.3) Then, the feature matrix was standardized (centered and scaled) and transformed with the Yeo-Johnson normalization via the *caret* R package (28). (C.4) Then, the data was partitioned using a 50:25:25 (train:test:discovery) split. (C.5) The discovery set was analyzed with the Boruta automated feature selection

method via the *Boruta* R package (29) to select the parsimonious feature set used to train the classifier. (C.6) The training set was then used to train a stacked ensemble classifier from the *caretEnsemble* R package (30), with the following architecture:

- The ensemble meta model is a Random Forest classifier and the 5 base models in the stack are:
    - Latent Dirichlet allocation
    - Recursive partitioning
    - Generalized linear model (logit)
    - K-nearest neighbors
    - Support vector machine (radial)
- 10-fold cross-validation repeated 5 times was implemented during training.

(C.7) Finally, the model was evaluated on the test set with the *confusionMatrix* function from the *caret* R package (28).

Sample classification (C)

Following model building, all samples were subsequently classified using the *classifySamples.R* script from the accompanying data generation repository (see *Code Accessibility*). This script primarily uses the *predictCondition* function from the *RLSeq* R package (see *Code Availability*). The procedure implements the following steps for each peak set in the reprocessed R-loop mapping data: (A) calculates the Fourier transform of the Z-score distribution, (B) reduces the dimensions to the engineered feature set (see *Quality model building*), (C) applies the preprocessing model to normalize these features (see *Quality model building*), and (D) implements the classifier to render a preliminary quality prediction (see *Quality model building*). Of note, the quality model alone does not provide the final prediction for each sample. Instead, the final prediction is "POS" (positive) only if all the following are true:

1. The RLFS Permutation test P value is significant ($p < .05$)
2. The Z-score distribution at 0bp is greater than 0 (see *R-loop forming sequences analysis*)
3. The Z-score distribution at 0bp is greater than the value at both -5000bp (start) and +5000bp (end) (see *R-loop forming sequences analysis*)
4. The quality model predicts a preliminary label of "POS".

These conditions were chosen to ensure that a positive prediction from the model (based on the distribution of the Z score) also met baseline criteria which indicate enrichment in typical permutation testing (27). These criteria may prevent spurious predictions when the model is applied to data types which it was not originally trained on.

Finally, predictions were rendered for every sample for which peaks could be calculated (656/693 samples).

Model validation (D)

To examine the internal and external validity of the predicted labels, several approaches were implemented: (A) RLFS consensus analysis, (B) genic feature enrichment analysis, and (C) sample correlation analysis. All 4 (prediction:label) conditions: POS:POS, POS:NEG, NEG:POS, NEG:NEG were compared in each approach.

RLFS consensus analysis

All samples for which peaks could be called by *macs3* (23) (37 samples excluded) were subjected to RLFS analysis (see *R-loop forming sequences analysis*). This produced a set of Z score distributions (one per sample). The data was scaled using min-max scaling, split by label:prediction group, and visualized using LOESS regression (span of 0.01) with standard error calculated.

Visualization of normalized coverage around MALAT1

This analysis involved four DRIP-Seq samples, "TCELL (Input)" (SRX2455193), "TCELL" (SRX2455189), "786-O (RNH)" (ERX3974965), and "786-O" (ERX3974964), in which both the "TCELL" samples were from the same study (SRP095885), and the "786-O" samples were also from the same study (ERP120322). Both "TCELL" samples and both "786-O" samples respectively were from the same study.

Feature enrichment analysis

Genic features were obtained via the *TxDb.Hsapiens.UCSC.hg38.knownGene* R package (31). The computationally predicted G4 quadruplex features were provided by *Chariker et al.* in their recent work (32). The experimentally determined G4 quadruplex features were determined by *Chambers et al.* via their previous chromatin immunoprecipitation (ChIP) sequencing study (GEO accession: GSE63874) (33).

All samples for which peaks could be called (37 samples excluded) were subjected to feature enrichment analysis for genic features and G4 quadruplex features. This analysis applied the following procedure to test the enrichment of peaks within these features: Using the *RLSeq* R package (1) peaks were randomly down sampled to a maximum of 10,000 peaks, (2) down sampled peaks were overlapped with genic features, (3) then, intersect statistics were calculated using Fisher's exact test via the *valr* R package (34), finally, (4) log2 odds ratios of each label:prediction group were compared using a Kruskal-Wallis test and pairwise Dunn pos-hoc tests with Bonferroni p value correction, all performed via the *rstatix* R package (35).

Correlation analysis

This analysis uses correlation of R-loop signal coverage around high-confidence R-loop sites (16) to assess sample-sample similarity. To prepare the analysis, the coverage track (bigWig) files for each sample were quantified around high-confidence sites (as first described by *Chedin et al.* (16)) via the *gsgCorr.R* script in the accompanying data generation repository (see *Code availability*). First, high-confidence R-loop sites were identified from ultra-long-read R-loop sequencing (SMRF-Seq) (36). Then, high-confidence sites were lifted from hg19 to hg38 and extended by 100kb bidirectionally. The sites were then binned into 1kb windows. Then, the coverage signal for every human R-loop mapping sample was summed over each bin to create a matrix of *bins[i] x samples[j]* containing the total signal of each sample within each bin [ij]. Finally, the *cor* function was used to calculate the Pearson correlation of the samples in the matrix. The results were visualized with the *ComplexHeatmap* R package (37).

R-loop consensus analysis

This analysis identified human R-loop consensus sites within catalytically dead RNase H1 (dRNH) and S9.6-based samples. To identify general R-loop consensus regions (RL regions), the union of dRNH and S9.6 consensus sites was calculated. To obtain these data, the following procedure was implemented: First, samples were prepared for R-loop consensus analysis using the *prepareConsensus.R* script from the accompanying data generation repository (see *Code Accessibility*). The following criteria were used to select samples for R-loop consensus analysis: (1) label of "POS", (2) prediction of "POS", (3) at least 5000 peaks called under a p adjusted value of 0.05, (4) is a human sample. We chose to enforce a label of "POS" as "NEG"-labeled are subjected to treatments which should prevent robust R-loop mapping and may, therefore, introduce unwanted technical variance to the consensus analysis. We chose samples with at least 5000 peaks due to the need for random peak down sampling to 5000 ranges.

227 samples met the full criteria for inclusion in the analysis of the 336 "POS"-labeled human samples. For each included sample, peaks were randomly shuffled and down sampled to 5000 ranges to ensure high peak-count samples would not dominate the analysis. Finally, samples were partitioned based on whether they used dRNH or S9.6-based R-loop mapping. Then, each down-sampled peak set was processed further using the *rlregions.smk* workflow (built on *snakemake*) from the accompanying data generation repository (see *Code Accessibility*).

For S9.6 and dRNH samples, the following were performed separately: (1) The peak sets were intersected with 10bp genomic windows and the number of peak overlaps per window was counted (using *bedtools* (38)). (2) The resulting windows and counts were converted to a bedGraph file (and its binary bigWig file format) and sorted using *bedtools* (38). (3) *macs3 bdgpeakcall* was then implemented with the options *-g 1000 -c x* where $x = \max\left(floor\left(.15 * n_{samples}\right), 5\right)$ and $n_{samples}$ is the number of samples provided (23). These steps yielded (A) consensus coverage tracks (bigWig) and (B) consensus R-loop peaks (narrowPeak) for both S9.6 and dRNH samples.

Then, consensus RL Regions were generated using the *peakUnion.R* script from the accompanying data generation repository (see *Code Availability*). This script found the union of the peaks in the dRNH and S9.6 narrowPeak files, along with calculating the percent of samples within each group in which each peak appeared (conservation score) and averaging these scores in cases where dRNH and S9.6 peaks overlapped to create a peak union. Finally, RL regions and associated metadata were finalized using the *finalizeRLRegions.R* script from the accompanying data generation repository (see *Code availability*). R-loops were filtered to remove any regions over 50kb in size, then every hg38 "POS"-labeled peak set was overlapped with the regions via the *bed_intersect* function from the *valr* R package (34) to yield a peak-RLRegion overlap matrix. This matrix was subsequently intersected with the RLBase sample metadata corresponding to each peak. Further details on additional metadata not used in this study are described in the *RLHub* R package reference (see *Code availability*).

Consensus site signal around genes

Consensus R-loop signal for dRNH and S9.6 was quantified around all hg38 ensemble genes using the *computeMatrix* function from the *deepTools* software program (24). The resulting data was scaled based on the number of samples in each group. The scaled signal was then scaled to a [0,1] range prior to plotting.

Consensus site overlap in genes

Consensus sites for dRNH and S9.6 were length-standardized by finding the highest signal within the peaks and extending 250bp bidirectionally to form 500bp ranges. These peaks were then overlapped with transcript features. Finally, a priority rule was enforced such that peaks which overlapped two annotation types would be assigned to the one with the highest priority. The priority order was “TSS”, “TTS”, “fiveUTR”, “threeUTR”, “Exon”, “Intron”, “Intergenic”.

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 accompanying data generation repository (see *Code availability*). *RLBase* sample alignment files (“BAM” format) were processed with *featureCounts* from the *Rsubread* R package to quantify the read counts from each within RL regions. Finally, the variance stabilizing transform (VST) was used to calculate normalized counts.

Differential abundance of RL regions

Differential abundance between dRNH and S9.6 samples was calculated by the following procedure: (1) the count matrix was subset to contain only the RL regions found by both dRNH and S9.6 peaks, (2) the *DESeq* function from the *DESeq2* R package (39) was applied to calculate differential abundance, (3) the *enrichR* R package (40) (and R interface to the enrichr web service (41)) was used to find the enrichment of gene ontology (GO) (42) terms within the genes overlapping differentially abundant RL regions.

Additionally, the *prcomp* function in R was used to calculate principal component analysis plots and the *EnhancedVolcano* R package (43) was used to generate the volcano plot. Finally, the pausing index for each gene (44) was calculated using the *getPauseIndices* function from the *BRGenomics* package (45). This calculation was performed using precision run-on (PRO) sequencing data in A549 cells downloaded from ENCODE (ENCFF275NOU and ENCFF979GYA).

Conservation analysis of RL regions

Conservation (percent of samples which detected an RL region) was ranked and bins were selected to partition RL regions by conservation percentiles. Binned RL regions were overlapped with genomic features and enrichment was calculated via the procedure described above (see *Consensus site overlap in genes*). Pathway enrichment was also performed via the procedure outline in the above section (*Differential abundance of RL regions*), with the only difference being that the ChEA (46), KEGG (47), and MSigDB (48) pathway databases were tested.

Results

Curation of public R-loop mapping samples

The data in this study was hand-curated from public data repositories, yielding 693 R-loop mapping samples from 65 different studies (**Table S1**). The data are technologically and biologically diverse, representing 58 different tissue types, 7 species, and 20 R-loop mapping modalities (**Table S1**, **Fig. S1A**). Moreover, the data includes both “POS” (expected to map R-loops) and “NEG” (not expected to map R-loops) samples (**Table S1**, **Fig. S1A**). Examples of “NEG” samples include “WKKD” (RNase H1 that is incapable of both DNA binding and R-loop resolution (12)) R-ChIP data, RNase H1-treated DRIP-Seq, and genomic “Input” samples. Examples of “POS” samples include “D210N” (catalytically inactive RNase H1) R-ChIP data and typical S9.6 DRIP-Seq data. The scale and diversity of these data made them suitable as a basis for building our quality methods and for our subsequent meta-analyses.

	Overall
n	693
Organism (UCSC ID) (%)	

Roundworm (ce11)	6 (0.9)
Zebrafish (danRer11	7 (1.0)
Fly (dm6)	30 (4.3)
Chicken (galGal6)	8 (1.2)
Human (hg38)	532 (76.8)
Mouse (mm10)	77 (11.1)
Yeast (sacCer3)	33 (4.8)
Mapping Mode (%)	
BisMapR	3 (0.4)
DREAM	12 (1.7)
DRIP	299 (43.1)
DRIP-RNA-Seq	2 (0.3)
DRIPc	40 (5.8)
DRIPc-HBD	2 (0.3)
DRIVE	1 (0.1)
DRNA	10 (1.4)
m6A-DIP	4 (0.6)
MapR	54 (7.8)
qDRIP	19 (2.7)
R-ChIP	54 (7.8)
RDIP	14 (2.0)
RLCnT-HBD	13 (1.9)
RLCnT-S96	16 (2.3)
RNH-CnR	5 (0.7)
RR-ChIP	5 (0.7)
S1-DRIP	12 (1.7)
sDRIP	49 (7.1)
ssDRIP	79 (11.4)
IP Type (%)	
dRNH	137 (19.8)
None	22 (3.2)
RNA-m6A	4 (0.6)
S9.6	530 (76.5)
Label = POS (%)	440 (63.5)

Table 2. Characteristics of the R-loop mapping samples profiled in this study. For each, the number and percentage of samples are shown. Organism (UCSC ID): The species from which data were derived along with the UCSC genomic ID used in alignment. Mapping Mode: the type of R-loop mapping used (see **Table S1** for greater detail). IP Type: the type of immunoprecipitation (IP) used during mapping (based on the mode) where “dRNH” indicates catalytically dead RNase H1, “None” indicates no IP, “RNA-m6A” indicates methylated RNA, and “S9.6” indicates use of the S9.6 antibody. “Label”: the binarized form of the “condition” metadata assigned by the authors (see Methods); “POS” indicates the sample is expected to map R-loops and “NEG” indicates the sample is not expected to map R-loops.

Following manual curation of R-loop mapping datasets, we developed a purpose-built R-loop pipeline tool called “RLPipes” (see *Code availability*) and applied it to reprocess all datasets, yielding peaks, coverage, and other processed files (see *Data availability*) (**Fig. 1A**).

The number of R-loop peaks varies within and between modalities

Previous work has demonstrated discrepancies in the number of peaks called between various R-loop mapping methods (13). However, to our knowledge, no study to-date has quantified the extent of these differences. From our analysis of the numbers of peaks called from various techniques, we found discrepancies between modalities spanning multiple orders of magnitude (**Fig. S1B**). We found that DNA:RNA *in vitro* enrichment (DRIVE) sequencing produced the fewest peaks (337). However, only one DRIVE-seq sample is publicly available at present (1). Following DRIVE-Seq, R-ChIP contains the fewest average peaks (~6 thousand). The greatest average number was generated by strand-specific DRIP (ssDRIP) sequencing (~314 thousand). Moreover, we observed variability among the number of peaks called even within mapping modalities (**Fig. S1B**).

Taken together, these findings suggest that the true locations of R-loops within a sample cannot be confidently assessed by any single modality, indicating that different modalities may preferentially detect different R-loops. The results also reveal technical variance between samples belonging to the same mapping modality, suggesting potential issues with sample quality control across studies.

R-loop forming sequences provide a suitable test of sample quality

Recent studies have indicated the presence of poor-quality R-loop mapping datasets within the literature (15, 16), but no tests of R-loop mapping accuracy have yet been proposed to our knowledge. To address this limitation, we developed a purpose-built quality control approach based on R-loop forming sequences (RLFS). RLFS are genomic regions which show favorability for R-loop formation (17, 49–52). We calculated RLFS across each genome and found that they agreed well with the R-loop mapping data we reprocessed (**Fig. 1B**). We then implemented permutation testing as a meta-analytic method of assessing the statistical enrichment of R-loop samples within RLFS. From this analysis, we found a strong enrichment of signal within RLFS for representative “POS” (positive) labeled samples and a lack of enrichment within “NEG” (negative) samples (**Fig. 1C**). These findings indicated the suitability of using RLFS as a basis for assessing R-loop mapping data quality.

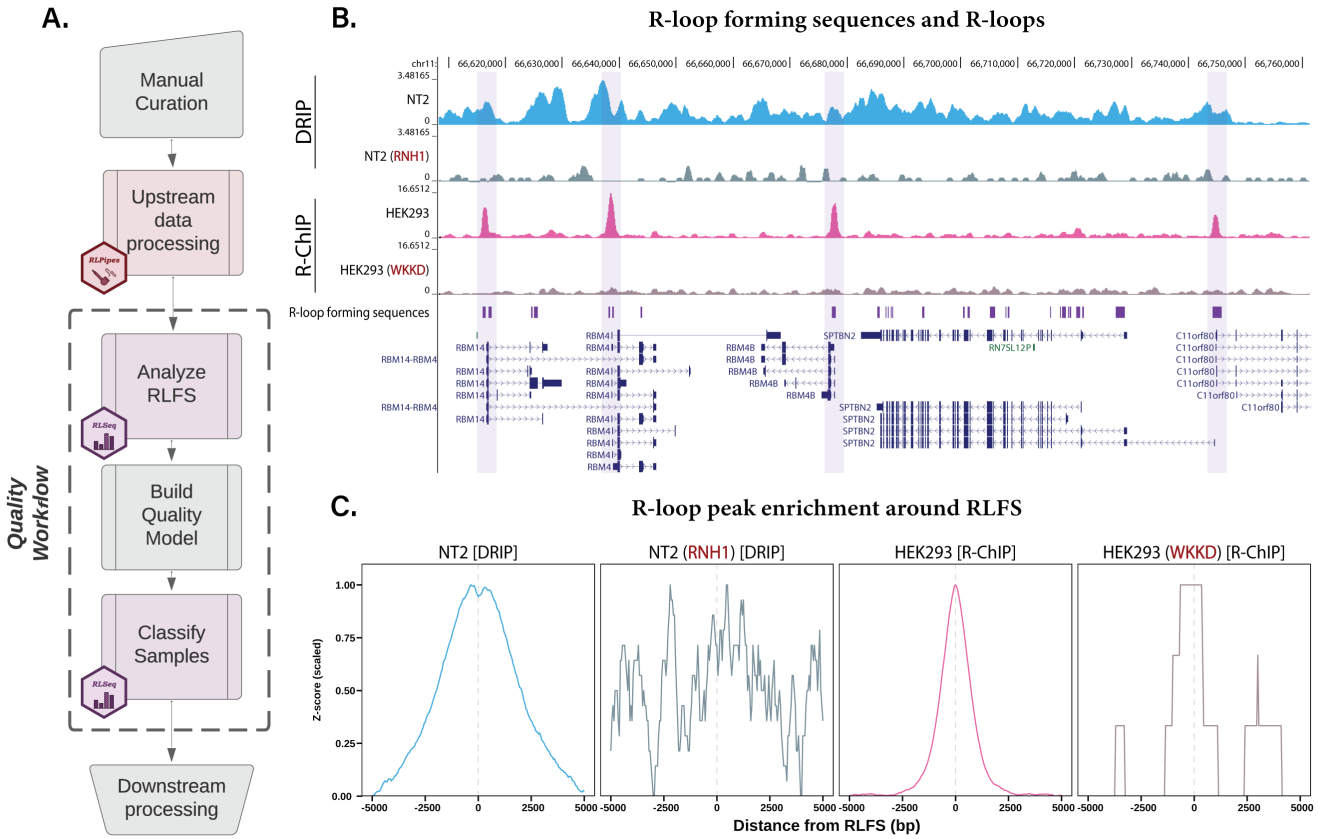

Figure 1. Reprocessing and standardization of R-loop mapping data. (A) The workflow used for reprocessed datasets along with icons indicating the software used at each step (see "Code availability"). (B) A Genome Browser view of four representative samples (one "POS" and one "NEG" each from DRIP and R-ChIP modes) along with R-loop forming sequences (RLFS). Negative label samples are indicated with red text (e.g., "RNH1"). The genome browser session for this visualization is also provided within this manuscript (see Data Availability). The SRA IDs of these samples are SRX1025894 ("NT2"), SRX1025896 ("NT2 (RNH1)"), SRX2683605 ("HEK293"), SRX2675009 ("HEK293 (WKKD)"; WKKD: mutant RNase H1 without catalytic or DNA binding capabilities). (C) Metaplots of the four samples in (B) showing the enrichment of their called peaks around RLFS. The Y axis is the Z score after min-max scaling.

Ensemble learning from RLFS provides accurate quality predictions

Evidence from our recent work and from *Chedin et al.* indicated that "POS"-labeled samples may sometimes fail to map R-loops (15, 16). To address this concern, we performed RLFS analysis for every sample in the present study for which peaks could be called (656 samples). As anticipated, we found some samples with a "POS" label, but which did not show expected enrichment around RLFS (**Fig. S2A**). This indicated the potential for using RLFS analysis to identify samples which fail to map R-loops as expected.

We reasoned that the p value from RLFS analysis would be sufficient for distinguishing sample quality. Unexpectedly, we found that, while many "POS" samples showed a favorable Z-score distribution and significant p value, there were also examples of false positives in which the p value was significant, but poorly representative of both the assigned label and of the Z score distribution (**Fig. S2B**). This indicated the need for a more sophisticated test of sample quality from the RLFS analysis results, prompting our development of an ensemble learning model which could analyze the "noisiness" of the Z score distribution to provide additional discriminatory capability.

The model was built using a semi-supervised approach that involves a human operator deciding whether to exclude samples from the training set based on the alignment between the RLFS Z-score plot and the sample label (**Fig. S2C-D**). Of the 656 samples which had sufficient read depth for peak calling, 135 (20.6%) were excluded from the analysis by the operator due to a mismatch between the sample label and the Z score distribution (see *Methods*). Having selected the samples to discard, the operator triggered the automated model building script, leading to a new model being trained (see *Methods*).

Nearly all features in the discovery set were automatically selected by Boruta for training (**Fig. S2E**). The ensemble learning model was then trained and evaluated internally via repeated cross validation, showing

average performance above 0.9 specificity, 0.8 sensitivity, and 0.9 receiver operator characteristics area under the curve (ROC) for all base models (**Fig. S2F**). The final performance of the model was evaluated on a test set which the model had never previously seen. On this data, the model displayed an accuracy of 0.9043 (CI: 0.8353-0.9513; No Information Rate (NIR): 0.6435; P-Value [Acc > NIR]: 1.263e-10) (**Fig. S2G**), demonstrating the high accuracy of the model.

Finally, we combined the machine learning model with three other assessments to develop our final quality model (see *Methods*). Briefly, a sample with a final prediction of "POS" must meet all the following criteria: (1) have a permutation testing p value below 0.05, (2) show a Z score above 0 at 0 bp in the distribution, (3) show a Z score at 0 bp which is greater than both the Z score at -5kb and +5kb, and (4) receive a preliminary prediction of "POS" from the machine learning model. The final "quality model" (criteria 1-4) was then used to predict a quality label for every reprocessed dataset in the present study (**Figure 2**).

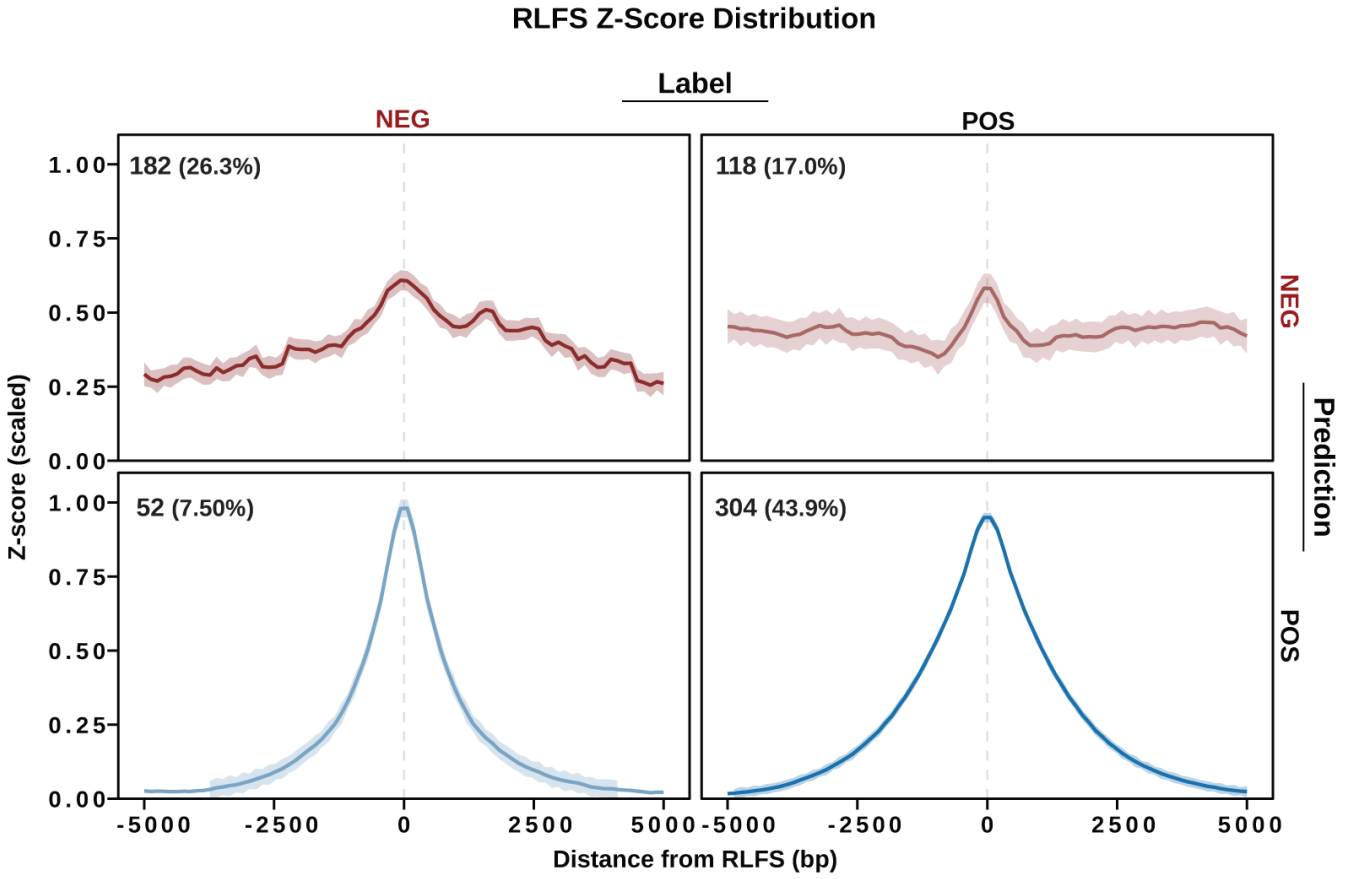

Figure 2: R-loop forming sequences analysis of prediction:label groups. Metaplots display the consensus signal among samples that were labeled "POS" (expected to map R-loops) or "NEG" (not expected to map R-loops) and samples which were predicted by the quality model to be "POS" or "NEG". The plots show the LOESS regression line along with a confidence interval based on the standard error from LOESS regression. The annotations show the number and percentage of samples which are included.

Validation of quality model reveals the prevalence of poor-quality R-loop mapping data

We then proceeded to assess the internal validity (consistency between predictions and RLFS analysis results from which predictions are made) and external validity (consistency between predictions and unrelated measures of mapping quality) of the quality model. All samples (even those which had been previously discarded) were tested with the model and assigned a classification (**Table S1**). Unexpectedly, we found a discrepancy between the assigned label (based on author-supplied metadata) and the predicted label (based on our quality model) for many samples. Of the 234 "NEG" labeled samples, 52 (22.22%) were classified as "POS", and, of the 422 "POS" labeled samples, 118 (27.96%) were classified as "NEG" (**Figure 2**). To better understand this unexpected result, we divided the data by all prediction:label combinations (NEG:NEG, NEG:POS, POS:NEG, POS:POS) and calculated the consensus Z score distribution for each group (**Figure 2, S3**). We demonstrate that the

classifications robustly represent the RLFS analysis of the data (**Figure 2, S3**). Coupled with the high accuracy of the ensemble model (**Fig. S2G**), these findings demonstrate that our quality model is internally valid.

We then proceeded to assess the external validity of our method through other quality approaches that did not involve RLFS analysis. The first approach was to visualize the normalized coverage tracks at sites of known R-loop formation. A representative depiction of four DRIP-Seq samples at a region near *MALAT1* (a non-coding RNA with constitutive R-loop formation (53)) shows that the model predictions distinguish between samples which display R-loop detection and those which do not (**Fig. 3A**). These results provided preliminary evidence for the external validity of our approach for determining whether a sample maps R-loops as expected.

We then implemented enrichment testing for genic features, such as exons and introns, across all human samples in each prediction:label group (see *Methods*). Given that R-loops are a byproduct of normal transcription, successful R-loop mapping should result in an enrichment of peaks within genic features, as has been shown previously (11, 12, 15). From our analysis, we found that samples that were predicted "NEG", regardless of their label, typically had low enrichment in all genic features (**Fig. 3B**). As predicted, we observed that the enrichment of genic features (with the exception of "Intron") was *not* significantly greater in NEG:POS (prediction:label) samples compared to NEG:NEG samples (**Fig. 3B**). Conversely, POS:NEG showed a significant increase in enrichment in all features when compared to NEG:NEG (**Fig. 3B**). Together, these results support the external validity of our model predictions.

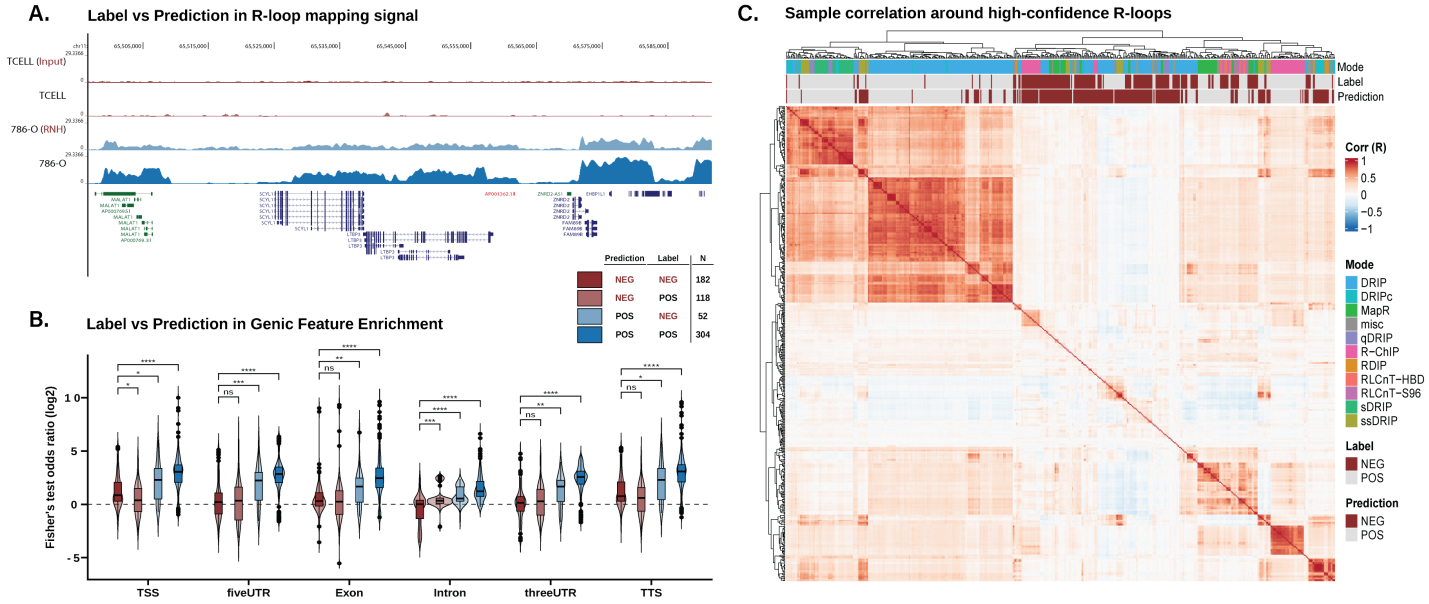

Figure 3. External validation of model predictions. (A) A Genome Browser image capture showing examples of DRIP-Sequencing datasets from two studies that exemplify all four possible model prediction versus sample label (prediction:label) combinations ("TCELL (Input)" [SRX2455193] - NEG:NEG; "TCELL" [SRX2455189] - NEG:POS; "786-O (RNH)" [ERX3974965] - POS:NEG; "786-O" [ERX3974964] - POS:POS). The genome browser session for this visualization is also provided within this manuscript (see Data Availability). (B) Violin/Box plots showing the distribution of Fisher's exact test odds ratios within all reprocessed R-loop mapping samples, split by prediction:label combination and by the genomic feature on which testing was performed. Significance was determined via the Kruskal-Wallis test followed by Dunn post-hoc with Bonferroni correction. **** - p < .0001; *** - p < .001; ** - p < .01; * - p < .05; ns – p >= .05. (C) A heatmap showing the Pearson correlation (R) of reprocessed R-loop mapping samples around high-confidence R-loop sites. Dendrograms and row order show hierarchical clustering of samples and annotations show the prediction, label, and mode of each sample.

Finally, we repeated the same analysis with both predicted and experimentally determined G4 quadruplexes (**Figure S4**). G4 quadruplexes are a non-B DNA structure which can form on the displaced, non-template strand of R-loops and may promote R-loop stability (54–56). From feature enrichment analysis of G4 quadruplexes, we found that "POS"-predicted samples were significantly more enriched than "NEG"-predicted samples, regardless of prior label (**Figure S4**). These findings further demonstrate the external validity of the model predictions and their applicability to the analysis of R-loop mapping results in a biological context.

Finally, we implemented sample-sample correlation analysis (see *Methods*) to evaluate the agreement between the results obtained from the model prediction and the sample-level similarity as determined through correlation (**Fig. 3C**). As expected, we found that "NEG"-predicted samples tended to cluster together and display low

correlation or no correlation with the majority of "POS"-predicted samples (**Fig. 3C**). Likewise, we found that predicted "POS" samples tended to cluster together, regardless of their original metadata label (**Fig. 3C**).

Taken together, these internal and external validations demonstrate that the method we present here is an accurate measure of R-loop mapping data quality. They also reveal the troubling extent of poor R-loop data within published datasets, indicating that the approach described here will be of great benefit to future studies. Of note, the method described here can be accessed via the *RLSeq* R package and the *RLBase* web application (see *Code availability*).

Definition of R-loop consensus regions

The locations and prevalence of R-loops genome-wide has been a subject of study since the first R-loop mapping modalities were described (1, 11). However, current estimates of where R-loops form and the percentage of the genome which they cover (3-13%) are derived from analysis of a relatively small number of samples and mapping technologies (11, 57, 58). As we demonstrate here, there is variation in the number of peaks called both between and within modalities (**Fig. S1B**), indicating that the true prevalence and locations of R-loops likely remain unclear.

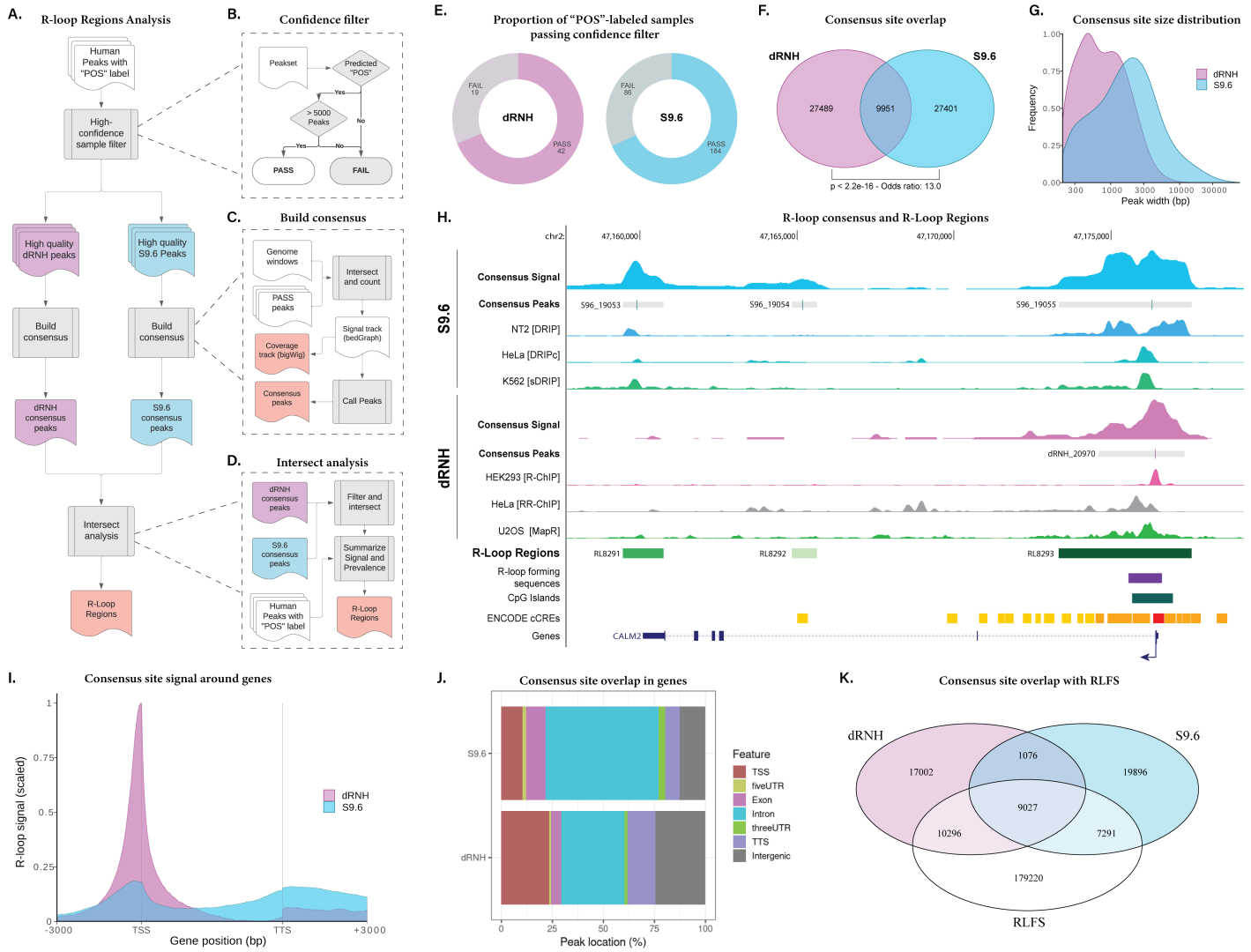

Figure 4. R-loop consensus analysis. (A) A flow diagram showing the workflow used for consensus analysis of R-loop sites in catalytically dead RNase H1 (dRNH) mapping samples and S9.6-based (S9.6) mapping samples. (B) A flow diagram showing the algorithm for selecting high-confidence samples to include in the analysis. (C) A flow diagram showing the workflow for building consensus R-loop signal and calling consensus peaks. (D) A flow diagram showing the workflow for intersect analysis, which yields R-Loop Regions. (E) Donut chart showing the proportion of "POS"-labeled dRNH and S9.6 samples passing the quality filters. (F) A Venn diagram showing the overlap of consensus sites derived from dRNH and S9.6-based mapping samples. P value and odds ratio from Fisher's exact test. (G) Density plot showing the frequency of consensus peak width derived from S9.6 and dRNH-based mapping samples. X axis is log10 scaled. (H) Genome browser view of the strongest non-repetitive R-loop region (at the CALM2 gene) with dRNH and S9.6 consensus signal, peaks, and signal tracks from representative S9.6 and dRNH R-loop mapping samples. The genome browser session for this visualization is also provided within this manuscript (see Data Availability). (I) Metaplot showing the R-loop consensus signal (scaled by

total number of samples in dRNH and S9.6 respectively) around genes. (J) Annotation plot showing the proportion of summitted dRNH and S9.6 consensus sites within various genomic features. (K) Venn diagram comparing dRNH and S9.6 consensus sites to R-loop forming sequences (RLFS).

To address this gap in knowledge, we analyzed high-confidence R-loop mapping samples to identify R-loop consensus regions (**Fig. 4A-D**), sites of robust R-loop formation across studies and modalities. Of note, our analysis workflow considers catalytically dead RNase H1 (dRNH) and S9.6 antibody (S9.6) R-loop mapping modalities independently prior to union analysis due to the hypothesized differences regarding the types of R-loops which these approaches map (14). The list of modalities belonging to the dRNH and S9.6 families is also provided within this manuscript (**Table S1**).

From the analysis workflow, we first defined a high confidence set of samples based on the label, prediction, and number of peaks in the sample (see *Methods*) (**Fig. 4B**). We found that 42 of 61 human "POS"-labeled dRNH samples and 184 of 270 human "POS"-labeled S9.6 samples passed the quality filters (**Fig. 4E**). The analysis proceeded as described (see *Methods*) and yielded consensus sites and signal for dRNH and S9.6 separately (**Fig. 4F**), along with a consensus set of "R-loop Regions" (RL regions) from an analysis of their union (**Table S2**) (see *Methods*). Of note, each region which is called as a "consensus site" was observed independently in at least 15% of all samples for both dRNH (6 out of 42 required) and S9.6 (27 out of 184 required) to be considered valid. This constraint helps ensure that the consensus sites likely represent true R-loop formation.

We found that dRNH samples yielded 39,840 consensus sites while S9.6 samples yielded 37,449 consensus sites, together covering 4.32% of the human genome. Moreover, we calculated the overlap of these sites and found 9951 shared sites (**Fig. 4F**). This indicated that, despite having clear differences, dRNH was finding many of the same R-loops as S9.6. We then proceeded to evaluate the size distribution of the consensus sites (**Fig. 4G**). This analysis revealed that S9.6 peaks tend to be much larger, recapitulating previous reports (12). We then proceeded to examine these sites at *CALM2*, a gene with one of the highest consensus scores (**Fig. 4H**). Interestingly, this visualization showed that, while both S9.6 and dRNH peaks were found around the TSS, only S9.6 was found in the gene body and around the TTS and PolyA regions (**Fig. 4H**). This finding led us to perform a meta-analysis of dRNH and S9.6 consensus sites and signal within genic features (**Fig. 4I-J**). From the results, we observed that while both S9.6 and dRNH show a preference for the TSS and TTS of genes, dRNH is more specific to these regions than S9.6 (**Fig. 4I**). Additionally, we calculated the proportion of overlaps of consensus R-loop sites for dRNH and S9.6-based samples within genic features, demonstrating that a greater proportion of dRNH consensus sites occur in promoter regions and intergenic regions, while a greater proportion of S9.6 sites occur in Introns (**Fig. 4J**). From repeating this analysis with regions split into "Shared", "dRNH-only", and "S9.6-only" groups, we observed a co-localization of "Shared" consensus sites within promoter and TTS regions (**Figure S5**). Moreover, from this analysis, we observed a over-representation of S9.6-only consensus sites in intronic regions and of dRNH-sites in intergenic regions (**Figure S5**). Finally, we compared the location of R-loop consensus sites and RLFS, finding that most shared consensus sites co-localize with RLFS (**Fig. 4K**).

Taken together, these findings indicated that consensus R-loop sites recapitulate expected R-loop biology. Moreover, they reveal strong differences between dRNH and S9.6 mapping approaches with respect to localization in genic features as predicted by *Castillo-Guzman and Chedin* (14). Finally, this analysis produced a union set of R-loop regions which can be used robustly for discrete analysis of R-loop mapping data (**Table S2**).

dRNH- and S9.6-specific R-loop regions display key biological differences with respect to promoter pausing and ribosomal gene programs

In their recent work, *Castillo-Guzman and Chedin* predicted that S9.6 and dRNH find different subsets of R-loops, "Class I" (R-loops resulting from promoter pausing) and "Class II" (R-loops resulting from transcriptional elongation) (14). They hypothesized that, while "Class I" R-loops are found by S9.6 and dRNH, "Class II" R-loops are more specific to S9.6 (14). The results of S9.6 and dRNH consensus sites described above support this hypothesis, but do not address promoter pausing or the biological programs impacted by the "Class I/II" distinction.

To address these questions, we proceeded to analyze differential R-loop abundance (see *Methods*) for each high-confidence sample within the 9951 R-loop regions (RL regions) found in both S9.6 and dRNH consensus sites (**Table S2**). We first performed principal component analysis (PCA) to visualize the variance within this dataset (**Fig. 5A**). This revealed that dRNH and S9.6 samples could still easily be separated even though the sites analyzed here represent consensus R-loop-forming sites (**Fig. 5A**). Differential abundance analysis led to

the identification of RL regions for which the S9.6-derived and dRNH-derived signal intensity were discordant (**Fig. 5B-C**, **Table S3**). We visualized the top differentially abundant RL region (RL51568), demonstrating the difference within this region between S9.6 and dRNH methods (**Fig. 5D**).

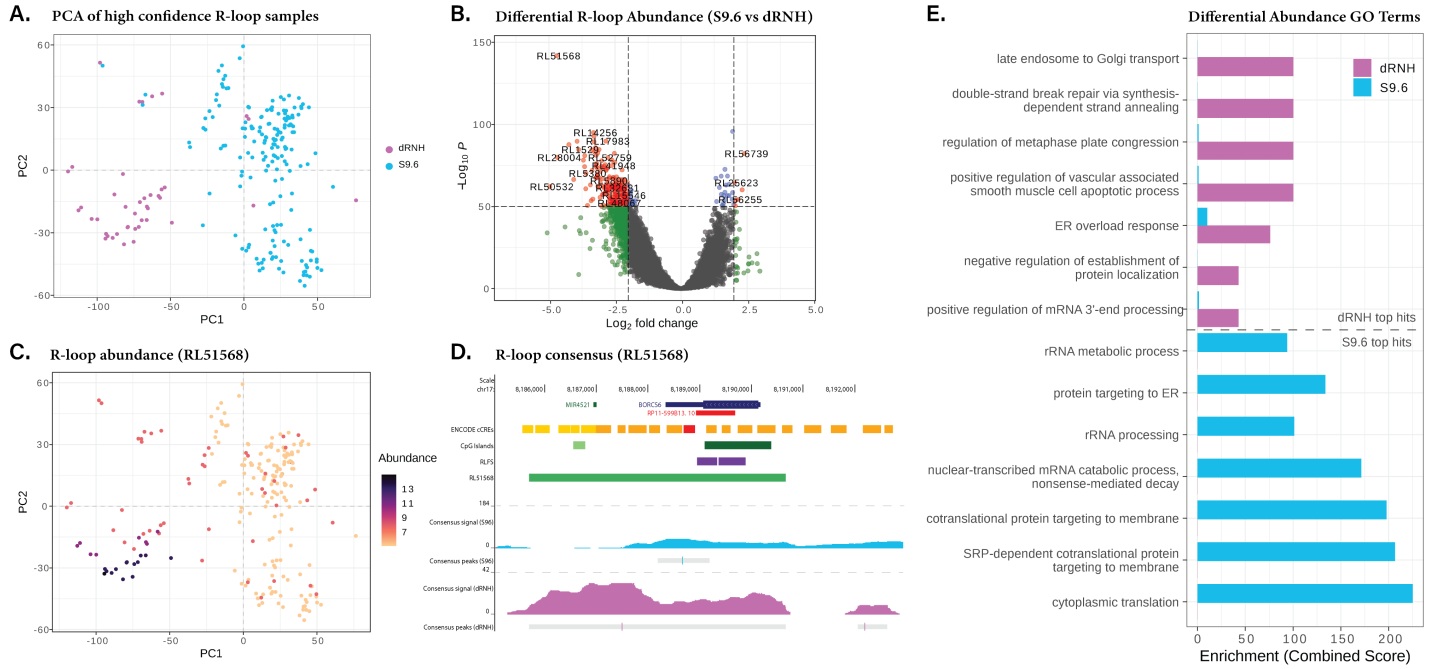

Figure 5. Differential abundance analysis of R-loop mapping signal. (A) A PCA plot showing the difference in R-loop profile between dRNH and S9.6-based approaches. (B) Volcano plot showing the top differentially abundant RL regions between dRNH and S9.6-based samples. (C) A PCA feature plot showing the top RL region ("RL51568"). (D) A genome browser view of S9.6 and dRNH consensus signal around RL51568. Coverage is scaled to the total number of high-confidence samples (S9.6: 184; dRNH: 42). (E) Top dRNH and S9.6-specific hits from GO term analysis of the genes associated with the top differentially abundant RL regions. Enrichment is measured by "Combined Score" such that higher indicates more strongly enriched.

To address the "Class I/II" hypothesis, we found the genes overlapping the dRNH-specific and S9.6-specific differential RL regions and calculated their pausing index (44). From the results, we observed that dRNH-specific genes show significantly greater pausing compared to S9.6-specific genes (2.25-fold) (**Fig. S6A**), supporting the validity of the "Class I/II" distinction.

To address the possibility that dRNH- and S9.6-specific R-loops occur in distinct biological programs, we found the genes overlapping the top 1000 in each and proceeded to perform GO term analysis (**Fig. 5E**). Unexpectedly, we found that R-loops in genes related to ribosomal programs were poorly detected in dRNH-based mapping methods but were strongly detected in S9.6-based mapping methods (**Fig. 5E**). To validate this finding, we examined the top 2000 dRNH and S9.6 consensus peaks and overlapped them with genes from the "KEGG Ribosome" pathway (as provided by the *msigdbr* R package (59)). Of 105 genes within the pathway, 58 were found in the top S9.6 consensus peaks, while only 13 were found in dRNH consensus peaks (**Fig. S6B**). In viewing the consensus signal around RPL17 (a ribosomal gene and one of the top differential RL regions), we observed a strong difference between S9.6 and dRNH signal (**Fig. S6C**). These findings further indicate that R-loops in many ribosomal genes are preferentially mapped by S9.6

Taken together, these results highlight the validity of the proposed "Class I/II" distinction as dRNH seems to show a strong preference for genes with more promoter pausing. Moreover, they demonstrate biological differences between dRNH-based and S9.6-based R-loop mapping in several gene programs, especially those involving ribosomal genes.

R-loop conservation analysis reveals biological differences between constitutive and variable R-loop regions

Finally, we addressed the question of where R-loop formation is consistent, regardless of biological context, and of where R-loop formation varies between tissue types. We determined R-loop region (RL region) *conservation percentages*, the proportions of high-confidence R-loop mapping samples that identify each RL region. Of note, this calculation was performed separately for dRNH and S9.6, and then averaged for sites found by both

techniques. From comparing dRNH and S9.6 conservation percentage distributions, we found that they tended to be relatively similar, with a noticeable increase in conservation levels within sites found by both dRNH and S9.6 (**Fig. S7A**). We then proceeded to examine the distribution of conservation percentages across all RL regions and assign them to bins which could be analyzed discretely (**Fig. 6A**).

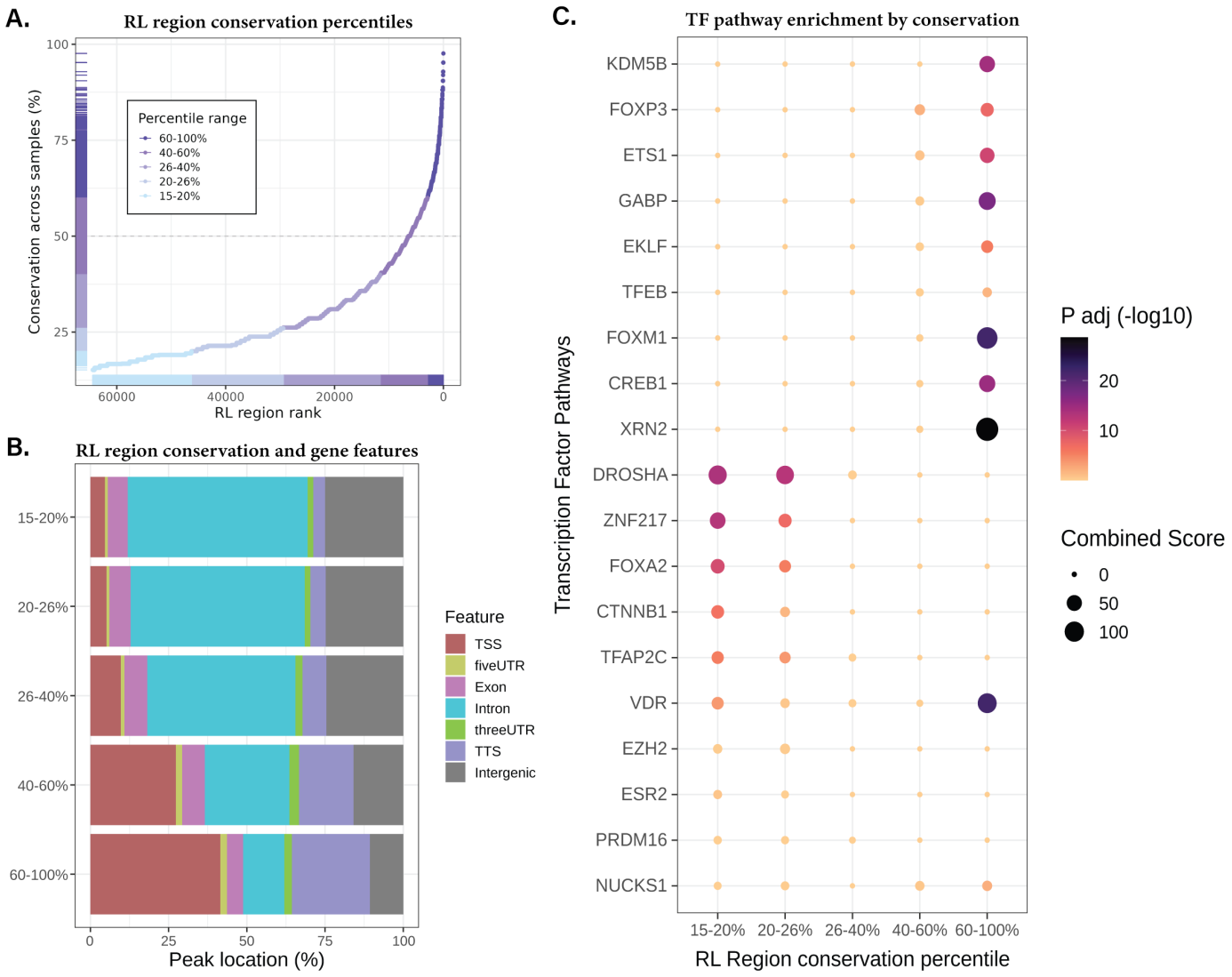

Figure 6. R-loop region conservation analysis. (A) Conservation rank plot showing all RL regions ordered by the percent of high-confidence samples in which they are found and binned by percentile ranges. (B) Annotation plot showing the gene features which overlap RL regions from various conservation percentile ranges. (C) Pathway enrichment plot showing the significance (via P adjusted value) and effect size (via Combined Score) of the "ChEA 2016" transcription factor pathway enrichment from the genes overlapping RL regions within each percentile.

Next, we examined the distribution of gene features overlapping the RL regions in each conservation bin (**Fig. 6B**). We found a strong association between high conservation bins and enrichment within the transcription start site and transcription termination site of genes (**Fig. 6B**). Moreover, we saw a stronger enrichment of intron and intergenic features in variable ("15-20%") RL regions (**Fig. 6B**). One possible explanation for this observation is that constitutive RL regions might associate with constitutively expressed "housekeeping genes" (60), while variable RL regions might overlap with more cell-type-specific genes. As predicted, we observed a greater enrichment of housekeeping genes in conserved bins (**Fig. S7B**). However, unexpectedly, the greatest enrichment was in the 40-60% conservation range (57 [-log10(padj)]) and not the 60-100% range (37.6 [-log10(padj)]) (**Fig. S7B**). This intriguing result led us to consider the possibility that conservation bins may represent biologically distinct R-loop programs that are partially independent from gene expression conservation.

To better understand these differences, we performed pathway enrichment on the genes overlapping each conservation bin (**Fig. 6C, S7C-D**). From an analysis of transcription factor binding pathways (**Fig. 6C**), we found

a divergence between constitutive and variable RL regions. The strongest hit from the constitutive regions was the "XRN2" target gene set. Notably, *XRN2* has been associated with R-loops in previous studies, particularly with respect to transcription termination (53). Conversely, the strongest hit from variable RL regions was the "DROSHA" target gene set. *DROSHA* is also a factor that has been previously studied in the context of R-loops (61). Moreover, we found *EZH2*, a core member of the polycomb repressive complex 2 (PRC2), was also associated with variable RL regions. Notably, PRC2 has been previously associated with R-loops in development-specific genes (62, 63). Curiously, we also observed the significant enrichment of the "VDR" target gene set in both constitutive and variable RL regions, but not in the bins between them (**Fig. 6C**). *VDR*, the vitamin D receptor, has not previously been associated with R-loop biology to our knowledge, further indicating the utility of this mode of analysis for revealing otherwise unknown associations with R-loop biology.

Having demonstrated the biological relevance of distinguishing between constitutive and variable RL regions, we proceeded to examine the other pathways enriched within these groups (**S7C-D**). Curiously, we noticed the enrichment of "G2-M Checkpoint" in constitutive RL regions and "Mitotic Spindle" in variable RL regions (**Fig. S7C**), indicating the potential of different classes of R-loops to act in the regulation of cell division. We also observed the enrichment of "Spliceosome" and "RNA transport" in constitutive RL regions (**Fig. S7D**). Finally, we found enrichment of the "Adherens junction" pathway in the variable RL regions (**Fig. S7D**).

Taken together, these results demonstrate that RL regions span a range of conservation levels, with many being strongly constitutive and others being more variable across samples. They also reveal that constitutive and variable R-loops associate with divergent genomic features and gene programs, indicating they may represent biologically meaningful subclasses of R-loops which require further exploration.

Discussion

The role of R-loops in physiology and pathology remains a topic of much study at present (3). Outstanding questions concern the types of R-loops which exist, their degree of conservation, their associations with epigenomic factors and processes, and the differences between various R-loop mapping approaches. Many of these questions cannot be faithfully answered by a single dataset, but instead must be gleaned through meta-analysis of hundreds of samples representing diverse biological conditions. In the present study, we mined and reprocessed 693 publicly available R-loop mapping datasets, developing the largest such resource to date. From these data, we developed a new and highly accurate method of quality control and implemented it to identify high-confidence R-loop mapping samples suitable for meta-analysis. From this analysis, we defined R-loop consensus sites and elucidated the differences between S9.6 and catalytically dead RNase H1 (dRNH) mapping approaches. Finally, we derived a set of "R-loop regions" (RL regions) from the union of dRNH and S9.6 samples and performed a conservation analysis which revealed the biological significance of constitutive and variable R-loops. The following sections summarize the key findings and provide additional context for their relevance to the field.

A robust quality model reveals the prevalence of poor-quality data in the literature

Within the R-loop mapping field, there are no suitable methods which can assess whether an individual sequencing sample has succeeded in robustly mapping R-loops. Instead, current studies rely only upon quality methods which are generic to all sequencing datasets, such as *FastQC* (12, 64). Previous meta-analyses of R-loop mapping datasets which implement quality assessments have relied upon bulk comparisons with previous data from the literature (15, 16), which can be confounded by systematic biases affecting large numbers of studies. From these limitations, and the previous evidence of inconsistent R-loop data quality (15, 16), we identified a clear need for a new R-loop mapping QC method that could benefit future researchers while also revealing the prevalence of high- and low-quality R-loop data within the field.

R-loop forming sequences (RLFS) are genomic regions which show favorability for R-loop formation (17, 49–52). They are computationally predicted from genomic sequence alone and show consistency with R-loop mapping data and with other computational prediction approaches, such as *skewr* (G or C skew prediction) (1, 2, 17). We calculated RLFS across each genome and found that they agreed well with the R-loop mapping data we reprocessed **(Figure 1)**, indicating they could be used as the basis for our quality control methodology. We then developed a quality model based on ensemble learning which uses the outputs from RLFS analysis to classify samples as "POS" (robust R-loop mapping) or "NEG" (poor R-loop mapping) (see *Methods*). This approach demonstrated excellent accuracy (**Fig. S2E-F**), along with both internal and external validity (**Figure 2, 3, S4**). Unexpectedly, we found many samples (24.5%) with mismatches between the assigned label and the predicted label from our model (**Figure 2**). Taken together, these results indicate the troubling extent of poor R-

loop data within published datasets, indicating that the approach described here will be of great benefit to future studies. Of note, the method described here can be accessed via the *RLSeq* R package and the *RLBase* web application (see *Code availability*).

R-loop consensus site analysis reveals the sites of robust R-loop formation

While it has been proposed that roughly 3-13% of the genome may contain R-loops, these findings were derived from an analysis of peaks from a limited number of S9.6-based R-loop mapping samples (11, 57, 58)(11). From our analysis, we found stark discrepancies in the number of peaks called between studies and between R-loop mapping modalities (**Fig. S1B**). These findings indicate that a multi-study and multi-modality consensus analysis is required to address the question of where R-loops form. From the meta-analysis we describe here, we identified robust sites of consensus R-loop formation which cover 4.32% of the human genome. Of note, this number may indicate a conservative estimate due to the quality controls we implemented. It will remain for future studies to incorporate more data to refine our estimate and calculate R-loop coverage in non-human genomes.

dRNH and S9.6 show differences in mapping Class I/II R-loops and distinctive biological programs

It was recently proposed by *Castillo-Guzman and Chedin* (14) that S9.6 and dRNH mapping approaches may map R-loops relevant to divergent transcriptional processes and programs. In particular, they propose a “Class I/II” distinction where “Class I” refers to R-loops which result from promoter-proximal pausing, and “Class II” refers to R-loops which result from transcriptional elongation. They hypothesize that dRNH may only find “Class I” R-loops whereas S9.6 may find both “Class I” and “Class II” R-loops. However, no bioinformatics analysis has yet demonstrated the existence of these R-loop classes genome wide.

From our analysis of dRNH and S9.6 consensus sites (**Figure 4, 5**), we found that S9.6-based R-loop mapping tended to produce larger peaks across the gene body while dRNH peaks tended to be smaller and more localized to promoter, TSS, or TTS sites as predicted by the “Class I/II” distinction. This was further confirmed by our analysis showing that dRNH-specific R-loop regions occur in genes which have a significantly higher pausing index (**Fig. S6A**). Altogether, these findings indicate the likely validity of the “Class I/II” hypothesis.

Given that S9.6 and dRNH appeared to map biologically distinct R-loops, we also decided to analyze the differences in gene programs found by each (**Fig. 5E**). Most striking was the observation that R-loop consensus sites found by S9.6, but not by dRNH, were strongly associated with pathways relevant for ribosome biogenesis and translation (**Fig. 5E, 6B-C**). At present, it is unclear why S9.6 seems to detect R-loops in ribosomal genes more efficiently or how this relates to the differences between “Class I” and “Class II” R-loops (14), indicating the need for further analysis by future studies.

Constitutive and variable R-loops occur in distinct biological pathways

Another outstanding question in the R-loop field relates to the dynamics of constitutive R-loops (those which are consistently abundant) and variable R-loops (those which form transiently). To address this question, we analyzed constitutive and variable sites of R-loop formation (**Figure 6, S7**). From our analysis, we found a strong relationship between RL region conservation and enrichment within TSS and TTS regions of genes. Because R-loops tend to form as a consequence of transcription in some genes, we decided to test whether high R-loop conservation was associated with “housekeeping genes”, genes which are thought to be constitutively expressed in most cell types (60). Curiously this analysis revealed that, while constitutive R-loops tend to be found in many housekeeping genes, these genes were not as strongly associated with the most highly conserved RL regions (**Fig. S7B**). This unexpected result may indicate that, while high R-loop conservation is likely associated with constitutive gene expression, there may be constitutive R-loops which form more independently from gene expression, possibly through trans-acting mechanisms. Additionally, from pathway enrichment we revealed a striking divergence in the biological programs and pathways enriched within constitutive and variable RL regions (**Fig. 6C, S7 C-D**). These findings indicate new and interesting directions for future R-loop studies to pursue.

Limitations

Hand-curation of metadata

The present study relies upon the accurate hand-curation of metadata from publicly available R-loop mapping studies. This process is time-consuming and may result in mistakes based on the care of the curator. It is also confounded by mistakes in the metadata being curated. We found, in some cases, that samples were mislabeled in the Gene Expression Omnibus (GEO), such that the labels for genomic input controls were switched with

positive R-loop mapping samples. In another instance, we observed that data which was publicly accessible was made private during our study. Finally, there was another instance in which the raw data files for specific samples in the Sequence Read Archive (SRA) were switched without notification on the GEO or SRA pages for the study. While there are ongoing initiatives to standardize the metadata in GEO and in the SRA (65), these initiatives are not expected to handle most of the issues described above.

To address these concerns, we have developed a data processing system which is adaptive to both new data and to changes in previous studies. The data processing workflow used in this study (see *Code availability*) checks the hand-curated data submitted by the user for marks that indicate samples which need to be re-assessed. Additionally, this workflow provides a convenient system for rebuilding the machine learning model used in quality assessment such that it remains adaptable to changes in the underlying data (described below). Finally, methods used in these analyses, such as R-loop region (RL region) abundance estimation and sample-level correlation analysis, provide visualizations that can be used to detect outliers that may be mislabeled. From these approaches, we expect the methods presented here to remain sensitive to mistakes in metadata and to remain adaptive to updates of previously profiled datasets and the addition of new data.

Quality model shortcomings

While the quality control method proposed in this study was demonstrated to be robust internally and externally (**Figure 2, 3, S2, S3, S4**), it may not be well-suited to assessing the R-loop mapping techniques which are developed in the future or those for which limited data is currently available. As of this writing, there are 23 different techniques for R-loop mapping, 14 of which were introduced since 2019, with seven introduced in 2021 alone (**Table S1**). Many of these techniques rely upon approaches that do not use either S9.6 or dRNH, such as m6A-DIP (66) and could pose unforeseen challenges for our quality model.

To address this limitation, we have developed an online learning scheme whereby an operator can, with minimal training and software experience, rebuild the model in a matter of minutes to include new datasets. To facilitate this process, we developed a user-friendly web-browser interface for interacting with the model-building software (**Fig. S2A-D**). This workflow offers a convenient and automatic model building scheme that will make it easy for us to add new data as it becomes available. This should ensure that the model will remain useful even on new R-loop mapping modalities.

Finally, it is possible that R-loop forming sequences (RLFS) may fail to capture certain subtypes of R-loops, especially those found in other species or from R-loop mapping modalities for which scant data is available. It was noted that data from species other than human and mouse often shows negative predictions (e.g., Fly) (**Table S1**) and some R-loop mapping modalities are more prone to receiving a negative prediction than others (e.g., DRIVE-Seq, m6A-DIP, etc) (**Fig. S2G**). At present, there is not enough data available in these modalities or species to evaluate whether the quality predictions exhibit a systemic bias. This highlights the importance for us to continue to rebuild the quality model as new data arises and repeat the analyses in this study at regular intervals when new data becomes available.

Consensus analysis

R-loop consensus was derived from 42 dRNH and 184 S9.6 high-confidence R-loop mapping samples. This represents all the available human data which passed our quality thresholds. However, it still indicates a significant imbalance between dRNH and S9.6, leading to potential inconsistencies in the accuracy of findings. For example, while there are only six studies from which the dRNH samples are derived, there are 27 for S9.6. This means that the dRNH set may be less representative of future studies, and some of the findings in this study may not be preserved upon incorporation of new datasets. While we are unable to address this issue today, we have made provisions to ensure that the process of incorporating new data is as convenient and automated as possible. Towards that end, we have developed a step-by-step protocol (see *Code Availability*) for recreating and extending the database we developed in this study. The protocol uses convenient shell scripts and automated *snakemake* workflows to reduce the time and effort necessary for adding new data. We expect this will make it easy for us to regularly update the database and, along with it, the robustness of the biological interpretation of these data.

Peak size between S9.6 and dRNH

Another caveat for the joint analysis of S9.6 and dRNH-derived data is that there are large discrepancies between the size of peaks, coverage regions, and consensus sites between them (**Fig. 1B-C, 4G**). S9.6-derived peaks, coverage, and consensus sites tend to be much wider than those from dRNH (**Fig. 1B-C, 4G**). Analyses for which this poses an immediate problem are those involving enrichment in genomic features when a feature

priority is used (**Fig. 4J, 6B**) as larger peaks may overlap with more genomic features, leading to spurious over-enrichment in top-priority features. Additionally, for analyses which involve finding genes that overlap with R-loop sites, larger peaks may lead to overlaps with many more genes than shorter peaks would find. To address this issue, we forced the peaks to a uniform width centered around peak "summits" (the areas of highest enrichment from peak calling). This technique allows us to mitigate the issues previously mentioned, but it may also have biases. S9.6 tends to have larger, more diffuse peaks. Particularly for noisier S9.6-based samples, the peak summits may not truly represent the point of highest enrichment and may be unable to fully account for the diffuse nature of the signal (**Fig. 4H**). Therefore, the summitting approach is used only sparingly for analyses which will directly suffer without it. Otherwise, the original peaks are preserved and reported without summits.

Bisulfite samples and stranded data analysis

Bisulfite samples were excluded from the analysis in this study due to the added complexity of processing these data and the low number of available samples (only 27 in total). However, these techniques may be useful in providing nucleotide-level R-loop mapping and may be more widely implemented in the future. Therefore, in future versions of the software used for this analysis, we will include methods for analyzing these data types.

Another limitation of this analysis is that it does not incorporate strand assignment for any of the strand-specific sequencing techniques included in the dataset. This is intentional as DRIP-Sequencing (which is not stranded) is still the most well-represented modality within the data and all analyses require conversion to unstranded data for compatibility with it. However, as stranded techniques gain in popularity and begin to outpace DRIP-Seq, we anticipate that future versions of the software used for processing these data will incorporate strand information. Therefore, we expect to eventually update the analyses presented in this study with stranded data to reveal higher-resolution insights into R-loop dynamics.

Conclusion

In this study, we reprocessed and standardized 693 public R-loop mapping samples, introducing a novel method of quality control to assess them. From the results, we demonstrate the widespread inconsistency of dataset quality within the literature and highlight the need for future studies to adopt our quality control approach. Moreover, from the high-confidence R-loop samples identified herein, we proceeded to define RL regions, sites of R-loop consensus across studies. From an analysis of these regions, we revealed the stark differences between the R-loops found by dRNH and S9.6-based techniques. Furthermore, we uncovered new biologically relevant subtypes of R-loops based on their conservation across samples. Taken together, this study introduces new methods and exciting future research questions for the R-loop mapping field.

Code Availability

The code used in this study is provided publicly on GitHub. The repository containing the workflow for recreating all upstream data analyses is provided here: https://github.com/Bishop-Laboratory/RLBase-data. The repository containing the code used to generate all figures in this manuscript is provided here: https://github.com/Bishop-Laboratory/RLoop-methods-and-analysis-2021. The software developed as part of the present study may be accessed here: https://gccri.bishop-lab.uthscsa.edu/rlsuite/. An interactive web application for exploring the data resource developed herein and analyzing user-supplied R-loop mapping peaks (*RLBase*) is available here: https://gccri.bishop-lab.uthscsa.edu/rlbase/.

Data Availability

All processed data files generated as part of this study are provided via *RLHub* and *RLBase*: https://gccri.bishop-lab.uthscsa.edu/rlsuite/. No new raw data was generated as part of this study. UCSC genome browser sessions shown in this study:

1. **Fig. 1B**: https://genome.ucsc.edu/s/millerh1%40livemail.uthscsa.edu/Fig_1B_rloops_rlfs
2. **Fig. 3A**: https://genome.ucsc.edu/s/millerh1%40livemail.uthscsa.edu/Fig_3A__Malat1
3. **Fig. 4H**: https://genome.ucsc.edu/s/millerh1%40livemail.uthscsa.edu/Fig_4H_rl_consensus
4. **Fig. 5D**: https://genome.ucsc.edu/s/millerh1%40livemail.uthscsa.edu/Fig_5D_RL51568
5. **Fig. S6B**: https://genome.ucsc.edu/s/millerh1%40livemail.uthscsa.edu/Fig_S6C_ribogenes

Acknowledgements

We want to thank all those who contributed to the development of the software used during this study, particularly the contributors to “RLSuite”. 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 the

Conflicts of interest

None to declare.

Funding

NIH/NCI [R01CA152063 and 1R01CA241554], CPRIT [RP150445] and SU2C-CRUK [RT6187] to A.J.R.B and Greehey Graduate Fellowship Award and NIH/NIA [F31AG072902] to H.E.M. NIH [R35 GM139549] to F.C. DOD [CDMRP PR181598] to K.S.

References

1. 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.

2. 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.

3. 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.

4. Hamperl,S., Bocek,M.J., Saldivar,J.C., Swigut,T. and Cimprich,K.A. (2017) Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell*, **170**, 774-786.e19.

5. Gan,W., Guan,Z., Liu,J., Gui,T., Shen,K., Manley,J.L. and Li,X. (2011) R-loop-mediated genomic instability is caused by impairment of replication fork progression. *Genes Dev*, **25**, 2041–2056.

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., 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.

8. Ngo,G.H.P., Grimstead,J.W. and Baird,D.M. (2021) UPF1 promotes the formation of R loops to stimulate DNA double-strand break repair. *Nat Commun*, **12**, 3849.

9. Arab,K., Karaulanov,E., Musheev,M., Trnka,P., Schäfer,A., Grummt,I. and Niehrs,C. (2019) GADD45A binds R-loops and recruits TET1 to CpG island promoters. *Nat Genet*, **51**, 217–223.

10. Li,Y., Song,Y., Xu,W., Li,Q., Wang,X., Li,K., Wang,J., Liu,Z., Velychko,S., Ye,R., *et al.* (2020) R-loops coordinate with SOX2 in regulating reprogramming to pluripotency. *Sci Adv*, **6**, eaba0777.

11. Sanz,L.A., Hartono,S.R., Lim,Y.W., Steyaert,S., Rajpurkar,A., Ginno,P.A., Xu,X. and Chédin,F. (2016) Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol Cell*, **63**, 167–178.

12. Chen,L., Chen,J.-Y., Zhang,X., Gu,Y., Xiao,R., Shao,C., Tang,P., Qian,H., Luo,D., Li,H., *et al.* (2017) R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. *Mol Cell*, **68**, 745-757.e5.

13. Yan,Q., Shields,E.J., Bonasio,R. and Sarma,K. (2019) Mapping Native R-Loops Genome-wide Using a Targeted Nuclease Approach. *Cell Rep*, **29**, 1369-1380.e5.

14. Castillo-Guzman,D. and Chédin,F. (2021) Defining R-loop classes and their contributions to genome instability. *DNA Repair (Amst)*, **106**, 103182.

15. 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.

16. Chédin,F., Hartono,S.R., Sanz,L.A. and Vanoosthuyse,V. (2021) Best practices for the visualization, mapping, and manipulation of R-loops. *EMBO J*, **40**, e106394.

17. 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**, W527-534.

18. Bushnell,B., Rood,J. and Singer,E. (2017) BBMerge - Accurate paired shotgun read merging via overlap. *PLoS One*, **12**, e0185056.

19. Chen,S., Zhou,Y., Chen,Y. and Gu,J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**, i884–i890.

20. Vasimuddin,Md., Misra,S., Li,H. and Aluru,S. (2019) Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*.pp. 314–324.

21. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R., and 1000 Genome Project Data Processing Subgroup (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

22. Faust,G.G. and Hall,I.M. (2014) SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*, **30**, 2503–2505.

23. Zhang,Y., Liu,T., Meyer,C.A., Eeckhoute,J., Johnson,D.S., Bernstein,B.E., Nusbaum,C., Myers,R.M., Brown,M., Li,W., *et al.* (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol*, **9**, R137.

24. Ramírez,F., Dündar,F., Diehl,S., Grüning,B.A. and Manke,T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res*, **42**, W187-191.

25. Lun,A.T.L. and Smyth,G.K. (2016) csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows. *Nucleic Acids Res*, **44**, e45.

26. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

27. Gel,B., Díez-Villanueva,A., Serra,E., Buschbeck,M., Peinado,M.A. and Malinverni,R. (2016) regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics*, **32**, 289–291.

28. Kuhn,M. (2008) Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, **28**, 1–26.

29. Kursa,M.B. and Rudnicki,W.R. (2010) Feature Selection with the Boruta Package. *Journal of Statistical Software*, **36**, 1–13.

30. Deane-Mayer,Z.A. and Knowles,J.E. (2019) caretEnsemble: Ensembles of Caret Models.

31. Bioconductor Package Maintainer,B.C.T. TxDb.Hsapiens.UCSC.hg38.knownGene.

32. Chariker,J.H., Miller,D.M. and Rouchka,E.C. (2016) Computational Analysis of G-Quadruplex Forming Sequences across Chromosomes Reveals High Density Patterns Near the Terminal Ends. *PLoS One*, **11**, e0165101.

33. Sahakyan,A.B., Chambers,V.S., Marsico,G., Santner,T., Di Antonio,M. and Balasubramanian,S. (2017) Machine learning model for sequence-driven DNA G-quadruplex formation. *Sci Rep*, **7**, 14535.

34. 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.

35. Kassambara,A. (2021) rstatix: Pipe-Friendly Framework for Basic Statistical Tests.

36. Malig,M., Hartono,S.R., Giafaglione,J.M., Sanz,L.A. and Chedin,F. (2020) Ultra-deep Coverage Single-molecule R-loop Footprinting Reveals Principles of R-loop Formation. *J Mol Biol*, **432**, 2271–2288.

37. Gu,Z., Eils,R. and Schlesner,M. (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, **32**, 2847–2849.

38. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

39. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*, **15**, 550.

40. Jawaid,W. (2021) enrichR: Provides an R Interface to 'Enrichr'.

41. Kuleshov,M.V., Jones,M.R., Rouillard,A.D., Fernandez,N.F., Duan,Q., Wang,Z., Koplev,S., Jenkins,S.L., Jagodnik,K.M., Lachmann,A., *et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res*, **44**, W90-97.

42. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25–29.

43. Blighe,K., Rana,S., Turkes,E., Ostendorf,B., Grioni,A. and Lewis,M. (2021) EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling Bioconductor version: Release (3.13).

44. Bunch,H., Zheng,X., Burkholder,A., Dillon,S.T., Motola,S., Birrane,G., Ebmeier,C.C., Levine,S., Fargo,D., Hu,G., *et al.* (2014) TRIM28 regulates RNA polymerase II promoter-proximal pausing and pause release. *Nat Struct Mol Biol*, **21**, 876–883.

45. DeBerardine,M. (2021) BRGenomics: Tools for the Efficient Analysis of High-Resolution Genomics Data Bioconductor version: Release (3.14).

46. Lachmann,A., Xu,H., Krishnan,J., Berger,S.I., Mazloom,A.R. and Ma'ayan,A. (2010) ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*, **26**, 2438–2444.

47. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*, **45**, D353–D361.

48. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S., *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, **102**, 15545–15550.

49. 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.

50. Lambo,S., Gröbner,S.N., Rausch,T., Waszak,S.M., Schmidt,C., Gorthi,A., Romero,J.C., Mauermann,M., Brabetz,S., Krausert,S., *et al.* (2019) The molecular landscape of ETMR at diagnosis and relapse. *Nature*, **576**, 274–280.

51. Heinäniemi,M., Vuorenmaa,T., Teppo,S., Kaikkonen,M.U., Bouvy-Liivrand,M., Mehtonen,J., Niskanen,H., Zachariadis,V., Laukkanen,S., Liuksiala,T., *et al.* (2016) Transcription-coupled genetic instability marks acute lymphoblastic leukemia structural variation hotspots. *Elife*, **5**, e13087.

52. Kuznetsov,V.A., Bondarenko,V., Wongsurawat,T., Yenamandra,S.P. and Jenjaroenpun,P. (2018) Toward predictive R-loop computational biology: genome-scale prediction of R-loops reveals their association with complex promoter structures, G-quadruplexes and transcriptionally active enhancers. *Nucleic Acids Res*, **46**, 7566–7585.

53. Villarreal,O.D., Mersaoui,S.Y., Yu,Z., Masson,J.-Y. and Richard,S. (2020) Genome-wide R-loop analysis defines unique roles for DDX5, XRN2, and PRMT5 in DNA/RNA hybrid resolution. *Life Sci Alliance*, **3**, e202000762.

54. Kumar,C., Batra,S., Griffith,J.D. and Remus,D. (2021) The interplay of RNA:DNA hybrid structure and G-quadruplexes determines the outcome of R-loop-replisome collisions. *Elife*, **10**, e72286.

55. Lee,C.-Y., McNerney,C., Ma,K., Zhao,W., Wang,A. and Myong,S. (2020) R-loop induced G-quadruplex in non-template promotes transcription by successive R-loop formation. *Nat Commun*, **11**, 3392.

56. Miglietta,G., Russo,M. and Capranico,G. (2020) G-quadruplex-R-loop interactions and the mechanism of anticancer G-quadruplex binders. *Nucleic Acids Res*, **48**, 11942–11957.

57. Wahba,L., Costantino,L., Tan,F.J., Zimmer,A. and Koshland,D. (2016) S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation. *Genes Dev*, **30**, 1327–1338.

58. Hartono,S.R., Malapert,A., Legros,P., Bernard,P., Chédin,F. and Vanoosthuyse,V. (2018) The Affinity of the S9.6 Antibody for Double-Stranded RNAs Impacts the Accurate Mapping of R-Loops in Fission Yeast. *J Mol Biol*, **430**, 272–284.

59. Dolgalev,I. (2021) msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format.

60. Hsiao,L.L., Dangond,F., Yoshida,T., Hong,R., Jensen,R.V., Misra,J., Dillon,W., Lee,K.F., Clark,K.E., Haverty,P., *et al.* (2001) A compendium of gene expression in normal human tissues. *Physiol Genomics*, **7**, 97–104.

61. Lu,W.-T., Hawley,B.R., Skalka,G.L., Baldock,R.A., Smith,E.M., Bader,A.S., Malewicz,M., Watts,F.Z., Wilczynska,A. and Bushell,M. (2018) Drosha drives the formation of DNA:RNA hybrids around DNA break sites to facilitate DNA repair. *Nat Commun*, **9**, 532.

62. Alecki,C., Chiwara,V., Sanz,L.A., Grau,D., Arias Pérez,O., Boulier,E.L., Armache,K.-J., Chédin,F. and Francis,N.J. (2020) RNA-DNA strand exchange by the Drosophila Polycomb complex PRC2. *Nat Commun*, **11**, 1781.

63. Skourti-Stathaki,K., Torlai Triglia,E., Warburton,M., Voigt,P., Bird,A. and Pombo,A. (2019) R-Loops Enhance Polycomb Repression at a Subset of Developmental Regulator Genes. *Mol Cell*, **73**, 930-945.e4.

64. Bayona-Feliu,A., Casas-Lamesa,A., Reina,O., Bernués,J. and Azorín,F. (2017) Linker histone H1 prevents R-loop accumulation and genome instability in heterochromatin. *Nat Commun*, **8**, 283.

65. Wang,Z., Lachmann,A. and Ma'ayan,A. (2019) Mining data and metadata from the gene expression omnibus. *Biophys Rev*, **11**, 103–110.

66. Abakir,A., Giles,T.C., Cristini,A., Foster,J.M., Dai,N., Starczak,M., Rubio-Roldan,A., Li,M., Eleftheriou,M., Crutchley,J., *et al.* (2020) N6-methyladenosine regulates the stability of RNA:DNA hybrids in human cells. *Nat Genet*, **52**, 48–55.