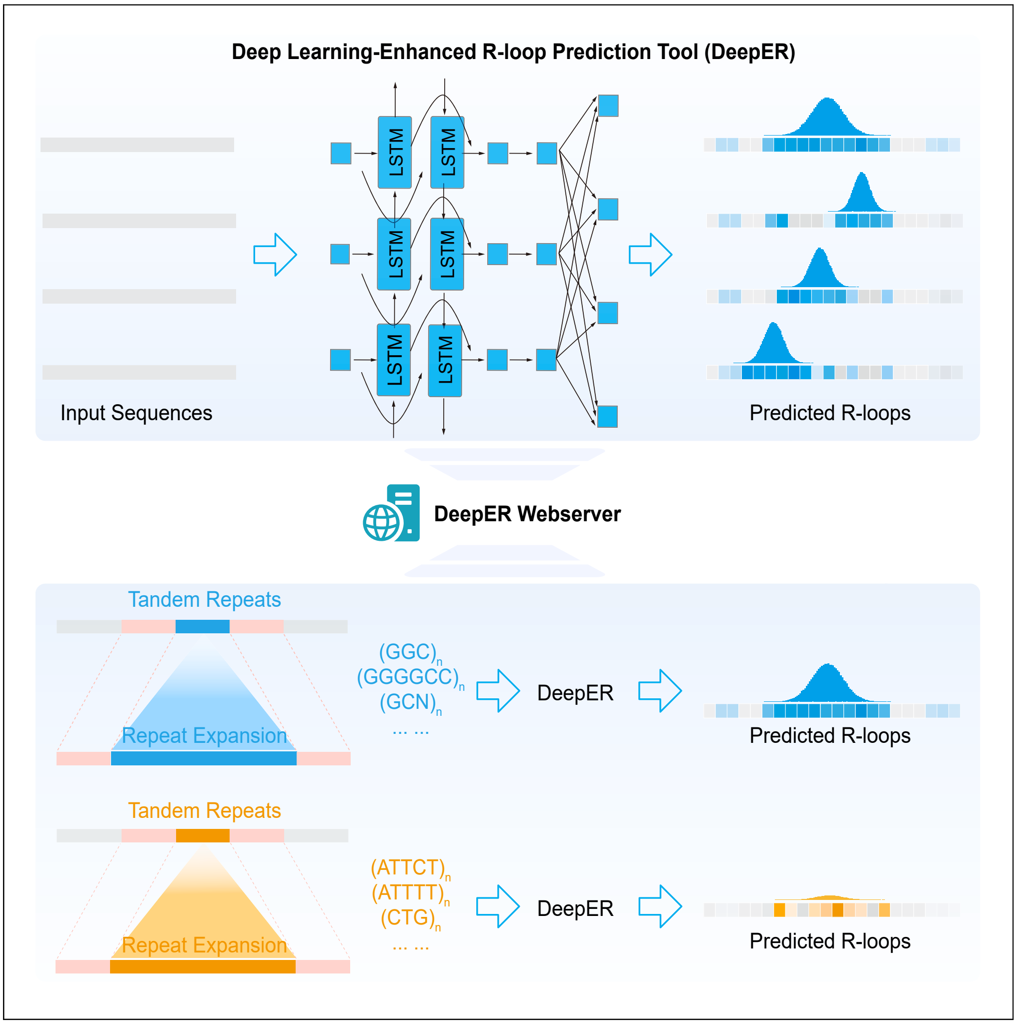
iScience

CellPress
OPEN ACCESS

Article

Deep learning-enhanced R-loop prediction provides mechanistic implications for repeat expansion diseases



Jiyun Hu, Zetong Xing, Hongbing Yang, ..., Yongcheng Pan, Lang He, Jia-Yu Chen

langhe@xupt.edu.cn (L.H.)
jiayuchen@nju.edu.cn (J.-Y.C.)

Highlights

DeepER, a deep learning R-loop prediction tool, demonstrates excellent performance

DeepER enables genome-wide annotation of R-loops and unveils key sequence features

DeepER predicts potential link of R-loops with certain repeat expansion diseases

Development of a web server for R-loop prediction enhances accessibility of DeepER

Hu et al., iScience *27*, 110584
August 16, 2024 © 2024 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.isci.2024.110584

iScience

CellPress
OPEN ACCESS

Article

Deep learning-enhanced R-loop prediction provides mechanistic implications for repeat expansion diseases

Jiyun Hu,[1,9] Zetong Xing,[1,9] Hongbing Yang,[1,9] Yongli Zhou,[1,9] Liufei Guo,[2] Xianhong Zhang,[3] Longsheng Xu,[1] Qiong Liu,[4] Jing Ye,[1] Xiaoming Zhong,[5] Jixin Wang,[6] Ruoyao Lin,[1] Erping Long,[6] Jiewei Jiang,[7] Liang Chen,[3] Yongcheng Pan,[4] Lang He,[2,*] and Jia-Yu Chen[1,8,10,*]

SUMMARY

R-loops play diverse functional roles, but controversial genomic localization of R-loops have emerged from experimental approaches, posing significant challenges for R-loop research. The development and application of an accurate computational tool for studying human R-loops remains an unmet need. Here, we introduce DeepER, a deep learning-enhanced R-loop prediction tool. DeepER showcases outstanding performance compared to existing tools, facilitating accurate genome-wide annotation of R-loops and a deeper understanding of the position- and context-dependent effects of nucleotide composition on R-loop formation. DeepER also unveils a strong association between certain tandem repeats and R-loop formation, opening a new avenue for understanding the mechanisms underlying some repeat expansion diseases. To facilitate broader utilization, we have developed a user-friendly web server as an integral component of R-loopBase. We anticipate that DeepER will find extensive applications in the field of R-loop research.

INTRODUCTION

R-loops, non-B nucleic acid structures composed of an RNA:DNA hybrid and a displaced single-stranded DNA, are key cellular regulators.[1–4] Dysregulated R-loops have been linked to various diseases, including neurodegenerative disorders, autoimmune diseases, and cancers.[5,6] Key to a better understanding of physiological and pathological roles of R-loops lies in the accurate detection of R-loops.

Two major types of experimental approaches have been developed to detect R-loops at a genome-wide level. The first relies on S9.6 antibody that can selectively recognize RNA:DNA hybrids, including DRIP-seq,[7] RDIP-seq,[8] DRIPc-seq,[9] bisDRIP-seq,[10] ssDRIP-seq,[11] qDRIP-seq,[12] *etc.* The second leverages a catalytically deficient but binding-competent RNase H mutant for R-loop enrichment, including DRIVE-seq,[7] R-ChIP,[13] RR-ChIP,[14] MapR,[15] bisMapR,[16] R-loop CUT&Tag,[17] *etc.* However, each approach exhibits inherent biases and may yield false-positive discoveries due to the utilization of distinct R-loop sensors and library construction strategies.[18–20] To date, there is no consensus yet regarding the number, size, and genomic distribution of detected R-loops.

Computational prediction of R-loop formation is an important complement to experimental approaches. R-loopBase[21] and RLBase[22] computationally deduced consensus R-loop regions; however, they relied on integrating existing R-loop mapping data. The RNA or the single-stranded DNA strand of R-loops is typically G-rich to ensure higher thermodynamic stability of the RNA:DNA hybrid[23] or permit G-quadruplex formation to facilitate RNA invasion.[24] Negative supercoiling can also promote R-loop formation.[25] QmRLFS-finder, R-loop tracker and other tools take advantage of the previous features to predict or characterize R-loop forming sequences.[7,26–29] These tools might, however, fail to detect R-loops associated with any unprecedented features. Deep learning has gained popularity in various genomic domains and achieved numerous successes,[30] having the potential to make accurate prediction and provide novel insights into unprecedented

[1]State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Chemistry and Biomedicine Innovation Center (ChemBIC), Department of Neurology at Nanjing Drum Tower Hospital, Nanjing University, Nanjing, Jiangsu 210023, China
[2]School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China
[3]RNA Institute, Hubei Key Laboratory of Cell Homeostasis, College of Life Sciences, Wuhan University, Wuhan, Hubei 430072, China
[4]Department of Neurology, Xiangya Hospital, Central South University, Changsha, Hunan 410008, China
[5]Center of Excellence for Leukemia Studies, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
[6]Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100005, China
[7]School of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an, Shaanxi 710121, China
[8]Nanchuang (Jiangsu) Institute of Chemistry and Health, Nanjing, Jiangsu 210023, China
[9]These authors contributed equally
[10]Lead contact
*Correspondence: langhe@xupt.edu.cn (L.H.), jiayuchen@nju.edu.cn (J.-Y.C.)
https://doi.org/10.1016/j.isci.2024.110584

features. The first deep learning-based tool in R-loop field, named deepRloopPre, was recently developed to predict R-loops in plants.[31] However, such a tool for studying R-loops in human has yet to be developed and applied.

Repeat expansion diseases are a group of genetic diseases resulting from expansion of short tandem repeats. Although advances in genome sequencing and genotyping have revealed ∼50 such diseases, the molecular mechanisms behind most of them remain unresolved.[32] Recognized pathomechanisms include genome instability, transcriptional repression, and the expression of aberrant RNA and protein products.[32] Noncanonical DNA secondary structures, facilitated by the pathological expansion of repeats, have emerged as pivotal players in disease pathology. For instance, the expansion of GAA repeats in the *FXN* gene of Friedreich's ataxia patients can lead to the formation of H-DNA, thereby suppressing transcription.[33] GGGGCC repeat expansions in *C9orf72* gene can form G-quadruplexes, triggering molecular cascades implicated in ALS/FTD pathogenesis.[34] Furthermore, expanded repeats have been implicated in promoting R-loop formation to induce transcriptional silencing or DNA damage responses.[6,34–37] However, it remains uncertain whether all types of repeats are associated with R-loop formation. *In vitro* transcription studies have suggested that some types of trinucleotide repeats are prone to R-loop formation,[38] but it is unclear whether these repeats induce R-loops in their native genomic contexts. We anticipate that a deep-learning model could potentially provide answers to these questions.

Here, we developed DeepER, both as standalone software and a web server, which allowed us to utilize deep learning techniques to accurately predict R-loop formation sites throughout the human genome. We identified crucial sequence features associated with R-loop formation, enabling us to gain mechanistic insights for some repeat expansion diseases.

RESULTS

Development of DeepER, a deep learning-enhanced R-loop prediction tool

We committed to developing a deep learning model that can make sequence-based prediction of R-loops. Considering that high-quality training data plays a crucial role in achieving optimal performance of a machine learning model, we prepared R-loop-positive and -negative datasets of higher reliability as follows (see STAR Methods): We utilized R-ChIP-mapped R-loops due to relatively higher accuracy and resolution,[18] and strand-specific signals. Specifically, we collected R-ChIP-mapped R-loop peaks conserved between K562 and HEK293 cells.[24] To further minimize R-ChIP-specific false positives, we kept only 3,204 peaks detected by at least one other R-loop mapping technology.[21] We randomly selected 5-kb-long intervals containing the previous R-loop peaks and surrounding R-loop-negative regions across the human genome. Ten intervals for each R-loop region were selected to enhance the model's robustness in handling R-loops at various positions relative to the 5-kb segment (Figure 1A). An approximately equal number of 5-kb-long intervals that did not contain R-loop peaks detected by any R-loop mapping technologies were randomly selected as additional R-loop-negative regions (Figure 1A).[21] Finally, we combined all selected intervals and allocated them into training, validation, and testing datasets at a ratio of 7:2:1. Their corresponding sequences were one-hot coded and used as input for training our deep learning model (Figure 1A).

Considering the critical roles of sequence contexts in R-loop formation, sequence-based R-loop prediction can be taken as a problem of long-term dependencies. Consequently, we built a deep-learning enhanced R-loop prediction tool (DeepER) using a residual BiLSTM model, known for its notable ability to address such challenges (Figure 1A and STAR Methods). One-hot coded input sequences underwent processing through one layer of BiLSTM followed by two layers of residual BiLSTM and a fully connected layer. After applying a sigmoid transition, DeepER generated probabilities of R-loop formation at the base level, ranging from 0 to 1. We then employed a sliding window approach with a window size of 200 bp and a step size of 10 bp to predict R-loop regions. A threshold value of 0.95 was then selected to optimize the F1 score, which represents the harmonic mean of the accuracy rate and recall rate. Adjacent windows with an average probability value ≥0.95 were merged to identify R-loop regions (Figure 1A). DeepER demonstrated good performance across all metrics for classification of R-loop-positive or -negative bases and regions (Figures 1B and 1C; Table S1). Of note, the AUROC values are 0.97 and 0.98 at the base and region levels, respectively, suggesting high discriminatory power and balanced performance (Figure S1A). The AUPRC values are 0.72 and 0.77 at the base and region levels, respectively, indicating adequate but improvable positive class identification (Figure S1B).

DeepER exhibited good generalization capability on an independent dataset. In this dataset, R-loop-positive regions were defined as subregions of DRIP-seq-mapped R-loop peaks that are sensitive to RNase H treatment,[18] which is a generally accepted "gold standard" for validating mapped R-loops. DeepER demonstrated comparable performance to that observed on testing data (Figure 1D), indicating its ability to effectively capture the underlying sequence features of R-loops and generalize well to new datasets.

Outstanding performance of DeepER compared to other R-loop prediction models

DeepER outperformed a series of alternative models trained using the same dataset. Feedforward neural networks failed to converge on the training dataset in most experiments. Extensive hyperparameter tuning was required to achieve convergence and satisfactory performance on the training data (Figures S2A and S2B). Despite these efforts, the performance of these models on the same testing dataset as DeepER was notably poor (Figure S2C), indicating overfitting and a failure to recognize the key features of R-loops, which rely on long-range dependencies. We also constructed a residual U-Net model for comparison by treating sequence-based R-loop prediction as a one-dimensional segmentation task. While the U-Net model successfully learned key sequence features and exhibited good generalization ability, its overall performance did not match that of DeepER (Figures S2D–S2G).

DeepER also showcased superior performance compared to other existing R-loop prediction tools. We conducted a comparative analysis between DeepER and deepRloopPre,[31] the first deep learning-based tool in the R-loop field, as well as R-loop tracker,[29] an efficient web-based implementation of the QmRLFS-finder algorithm.[26] By referring to the original literature of these three tools,[29,31] we found that

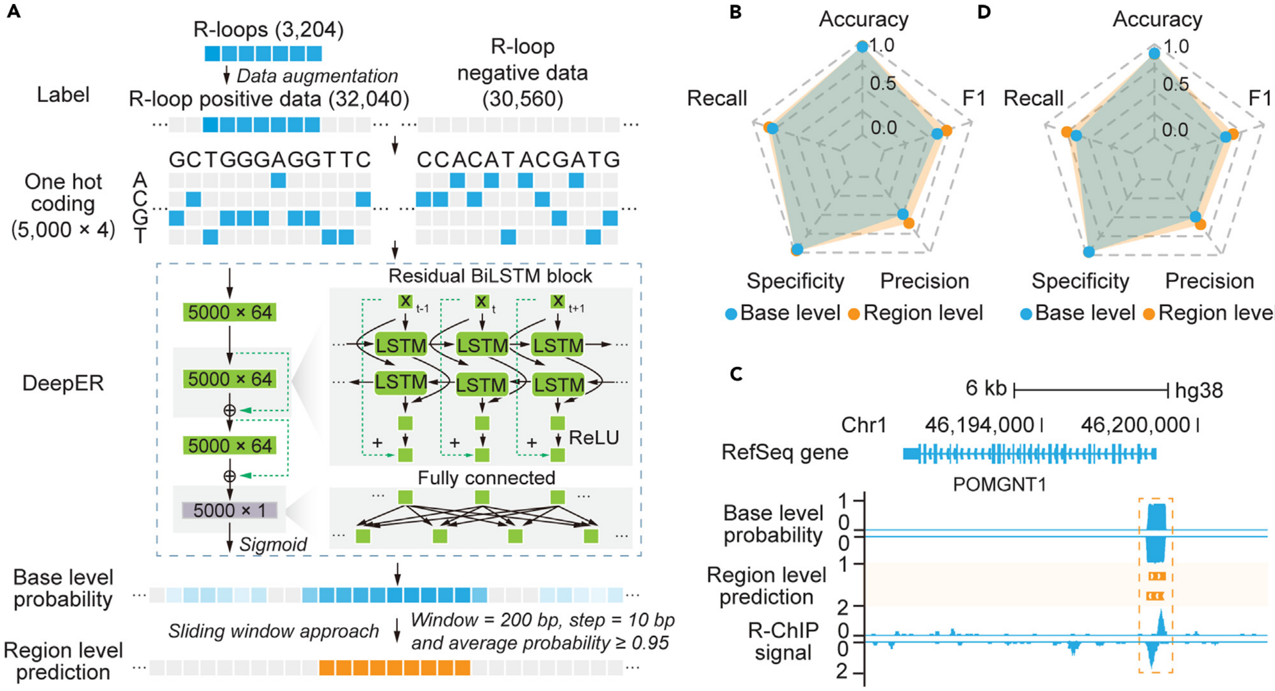


Figure 1. The framework and performance of DeepER

(A) R-loop-positive and -negative bases are labeled as 1 (blue) and 0 (gray), respectively. DeepER receives one-hot coding of R-loop-positive and -negative sequences as input, uses BiLSTM blocks with residual layers as basic framework to predict base-level probability value of input sequence. Adjacent sliding windows (window size = 200 bp and step size = 10 bp) with the average probability value ≥0.95 are merged and defined as predicted R-loop regions.

(B) DeepER's performance evaluated with the testing dataset.

(C) Predicted base-level probability values and R-loop regions in comparison with R-ChIP-mapped R-loop signals (RPM) at POMGNT1 gene locus.

(D) DeepER's performance evaluated with an independent dataset of experimentally verified R-loops.

DeepER demonstrates better recall, F1 score, accuracy and specificity than the other two tools (Table S1). Only the precision (0.61) is lower than R-loop tracker (0.73), but still comparable with deepRloopPre (0.64) (Table S1). Although the previous comparison is based on different datasets, it still provides circumstantial support for the superior performance of DeepER. Furthermore, we applied all three tools to predict R-loops across the human genome for direct comparison. Currently, there is still a lack of a gold standard dataset for R-loop regions across the human genome. Here, we used consensus R-loop regions deduced by RLBase[39] or R-loopBase[21] from existing R-loop mapping data to evaluate the performance of all three tools. Clearly, DeepER outperformed the other two tools in 5 out of 7 metrics, with the remaining metrics being comparable to those of the best tools (Figure S3; Table S2).

However, when running on CPU, DeepER required about five times as much time to complete the prediction compared to the other tools, and it also incurred a larger memory cost. This is attributed to the inclusion of two additional layers of BiLSTM and approximately 5-fold more parameters than DeepRloopPre (235,650 vs. 41,550), which are necessary to achieve better prediction performance (Figure 1B; Table S1). Importantly, these resource requirements significantly decreased once DeepER was executed on GPU (Table S3). We believe that, as a trade-off for the improved prediction results, the relatively longer processing time and increased memory usage of DeepER are acceptable.

Whole-genome annotation of R-loop-forming sequences with DeepER

The outstanding capabilities of DeepER enabled us to make accurate predictions of R-loop regions across the entire human genome. In total, 79,626 R-loop regions were predicted. The sizes of predicted R-loop regions ranged from 200 bp to 3,050 bp (mean = 538 bp and median = 400 bp) (Figure 2A), slightly longer than R-loops in the training dataset (Figure S4A). Although R-loops in the training dataset were almost exclusively located at transcription start sites (TSS) (Figure S4B), 42.4% of predicted R-loops were located at transcription termination sites (TTS), gene body and intergenic regions (Figure 2B), suggesting DeepER's capability to learn features from the training dataset, and react properly to previously unseen, new data.

The R-loop regions predicted by DeepER were in good agreement with those identified by other experimental or computational methods. Around 45% of the R-loops predicted by DeepER were also detected by QmRLFS-finder algorithm, which searched for G-rich sequences based on a pattern-based rule[26] (Figure 2C). The remaining 55% DeepER-specific R-loops likely represented R-loops characterized by other sequence features missed out by QmRLFS-finder. About 80% of the predicted R-loops could be detected by at least one R-loop-mapping technology, and these were classified as class I R-loops (Figure 2D). As shown in Figure 2E, one class I R-loop at BCL10 promoter that was

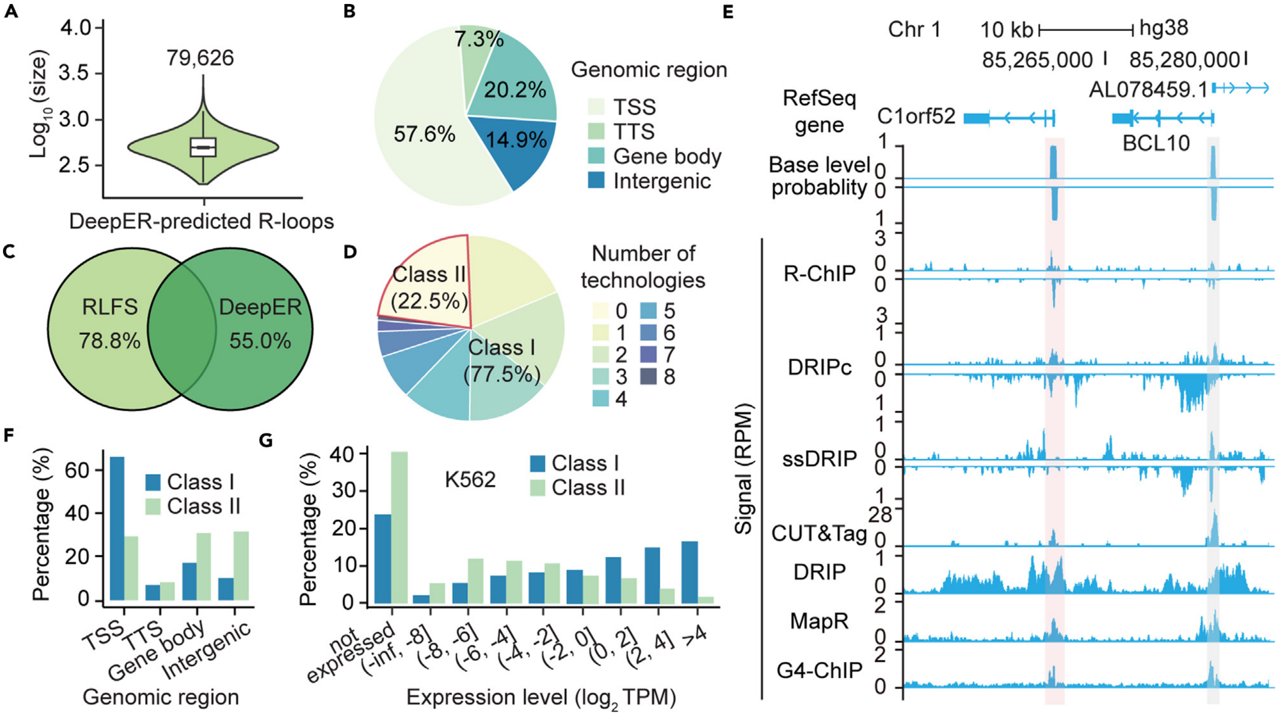


Figure 2. Genome-wide R-loop prediction by DeepER

(A and B) The size (A) and genomic distribution (B) of R-loops predicted by DeepER.

(C) Comparison of R-loops predicted with DeepER and QmRLFS-finder. Numbers indicate percentages of method-specific R-loops.

(D) Fractions of predicted R-loops detected with different numbers of R-loop mapping technologies. R-loops detected by one or more technologies are defined as class I R-loops, and the others as class II R-loops.

(E) Examples of predicted R-loops in reference to R-loop peaks detected by different R-loop detection technologies. Highlighted in the pink and gray rectangles are R-loop peaks included and not included in the training dataset, respectively.

(F) Genomic distribution of class I and class II R-loops.

(G) Class I and class II R-loop numbers as a function of gene expression levels.

not detected by R-ChIP, and therefore not included in the training dataset, was detectable by many other R-loop mapping technologies. The remaining ∼20% of predicted R-loops, which were not supported by any R-loop mapping technologies, were classified as class II R-loop regions. DeepER was trained with sequence information only, it likely failed to identify R-loop forming characteristics not encoded in sequences. We hypothesized that the DNA sequences of these regions did contain features that could promote R-loop formation; however, there was probably no complementary RNA generated to invade the DNA for R-loop formation. Consistent with our prediction, more class II R-loops were found in intergenic regions than class I R-loops (Figure 2F). Furthermore, class II R-loops were generally associated with non-expressed or lowly expressed genes in different cell lines (see Figures 2G and S4C–S4E).

Position- and context-dependent effects of G nucleotides on R-loop formation

The sequential nature of input data posed challenges to feature importance analysis for BiLSTM model. Here, we explored the sequence features important for DeepER-predicted R-loops using a permutation-based method (Figure 3A). Briefly, we introduced one single point mutation at a random position within each R-loop region and re-evaluated the formation probability with DeepER. Mutations leading to probabilities falling below the threshold for R-loop region prediction were classified as R-loop-disrupting mutations. In contrast, mutations that preserved the integrity of the R-loops, referred to as R-loop-preserving mutations, were taken as control. We then compared these two mutation classes to dissect the crucial sequence features governing R-loop formation. About 11.5% of the initial R-loop regions exhibited disruption, as exemplified in Figure 3B. Stronger disrupting effects were observed for R-loop-disrupting mutations than R-loop-preserving mutations (Figure S5A). Notably, R-loop-disrupting mutations exerted a discernible impact on a long genomic stretch surrounding the site of mutation (Figure 3C). This observation aligns with the concept that the BiLSTM model utilizes information from flanking regions for R-loop prediction, enabling further investigation of sequence features.

We unveiled the significance of position-aware sequence composition for R-loop formation. Notably, R-loop-disrupting mutations displayed a strong bias toward G bases, mainly manifesting as G-to-C and G-to-T mutations (Figure 3D). Both mutation types likely contribute to reduced thermodynamic stability of RNA:DNA hybrids, stemming from decreased GC-skew levels[23] or hydrogen bond counts. Furthermore, we revealed a notable enrichment of G bases at both upstream and downstream regions of the R-loop center (Figure 3E). This

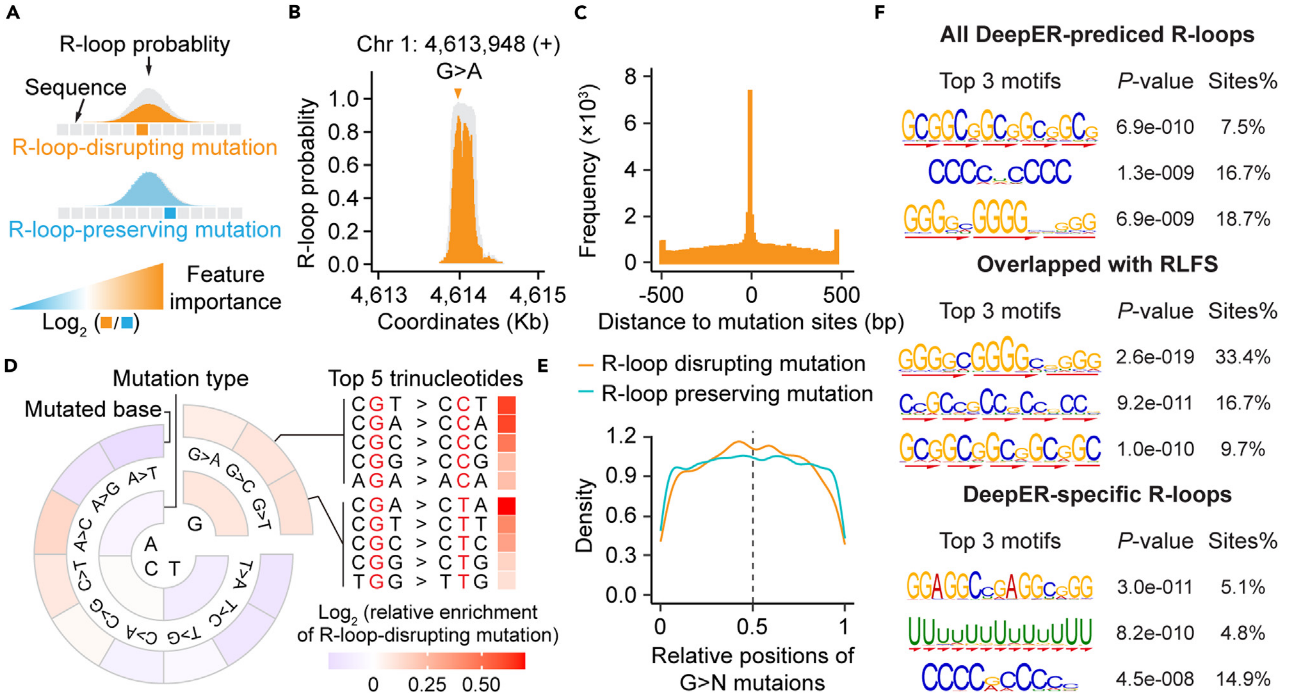


Figure 3. Feature importance interpretation of DeepER

(A) Schematic representation of two classes of mutations that will disrupt (orange) or preserve (light blue) the DeepER-predicted R-loops.

(B) Base-level R-loop formation probabilities at a representative genomic locus before (gray) and after (orange) the introduction of an R-loop-disrupting G>A mutation.

(C) Distribution of sites showing maximum negative change of probability. Zero point represents the mutation site.

(D) Relative enrichments of mutated bases, mutation types and 3-mers defined as the ratio of R-loop-disrupting mutations to R-loop-preserving mutations.

(E) Distribution of R-loop-disrupting and R-loop-preserving G > N (N = A, C, and T) mutations across R-loop regions. The R-loop center is indicated by a dashed line.

(F) Top 3 enriched sequence motifs of all DeepER-predicted R-loops (top), R-loops that were co-detected by QmRLFS-finder (middle) and DeepER-specific R-loops. Short tandem repeats are indicated with underlines.

observation aligns with the higher thermodynamic stability requisite for R-loop initiation and extension proposed previously.[26] Intriguingly, although A bases did not exhibit overrepresentation in R-loop-disrupting mutations, the A-to-C mutation type displayed specific enrichment, generally consistent with previous findings linking purine-rich[24] and AT-skewed[11] sequences to R-loop formation. Analyzing their distribution across R-loop regions unveiled mutated A bases predominantly positioned downstream of R-loop centers (Figure S5B). Our findings offer insights into the intricate molecular mechanisms underpinning R-loop formation.

We further noted context-dependent importance of G nucleotides. In line with the role of G-quadruplex structures in promoting R-loop formation,[40] we revealed an enrichment of GGG (Figure 3F, top and middle panels), the basic unit of G-quadruplexes, and a significant association between DeepER-predicted R-loops and G-quadruplex formation (Figures S5C and S5D). Interestingly, G nucleotides downstream of C nucleotides were in general overrepresented in R-loop disrupting mutations (Figure 3D). Consistently, the top-ranked sequence motifs for DeepER-predicted R-loops all exhibited rich GC content, particularly pronounced for those R-loops co-detected by QmRLFS-finder (Figure 3F, top and middle panels). Interestingly, these sequence motifs were strongly associated with GCG, CGC, and GGGGC tandem repeats (Figure 3F). This observation underscores the potential connection between R-loop formation and the occurrence of some repeat expansion diseases (see below).

DeepER specifically identified R-loops characterized by GA-rich sequences (Figure 3F, bottom panel), which aligns with the roles of purine-rich[24] and AT-skewed[11] sequences in R-loop formation. Surprisingly, poly-C and poly-U sequences were also found to be enriched in DeepER-specific R-loops (Figure 3F). These sequences might originate from R-loop formation on the opposite strand, considering the prevalence of antisense R-loops.[11]

DeepER predicts the potential link of R-loop formation with some repeat expansion diseases

Motivated by the observed association between tandem repeats and R-loop formation, we employed DeepER to analyze all documented tandem repeats associated with repeat expansion diseases (Table S4). Repeat numbers were increased up to 200 copies beyond the reported minimum pathogenic repeat number, or decreased to the repeat number of the reference genome or 200 copies less than the minimum

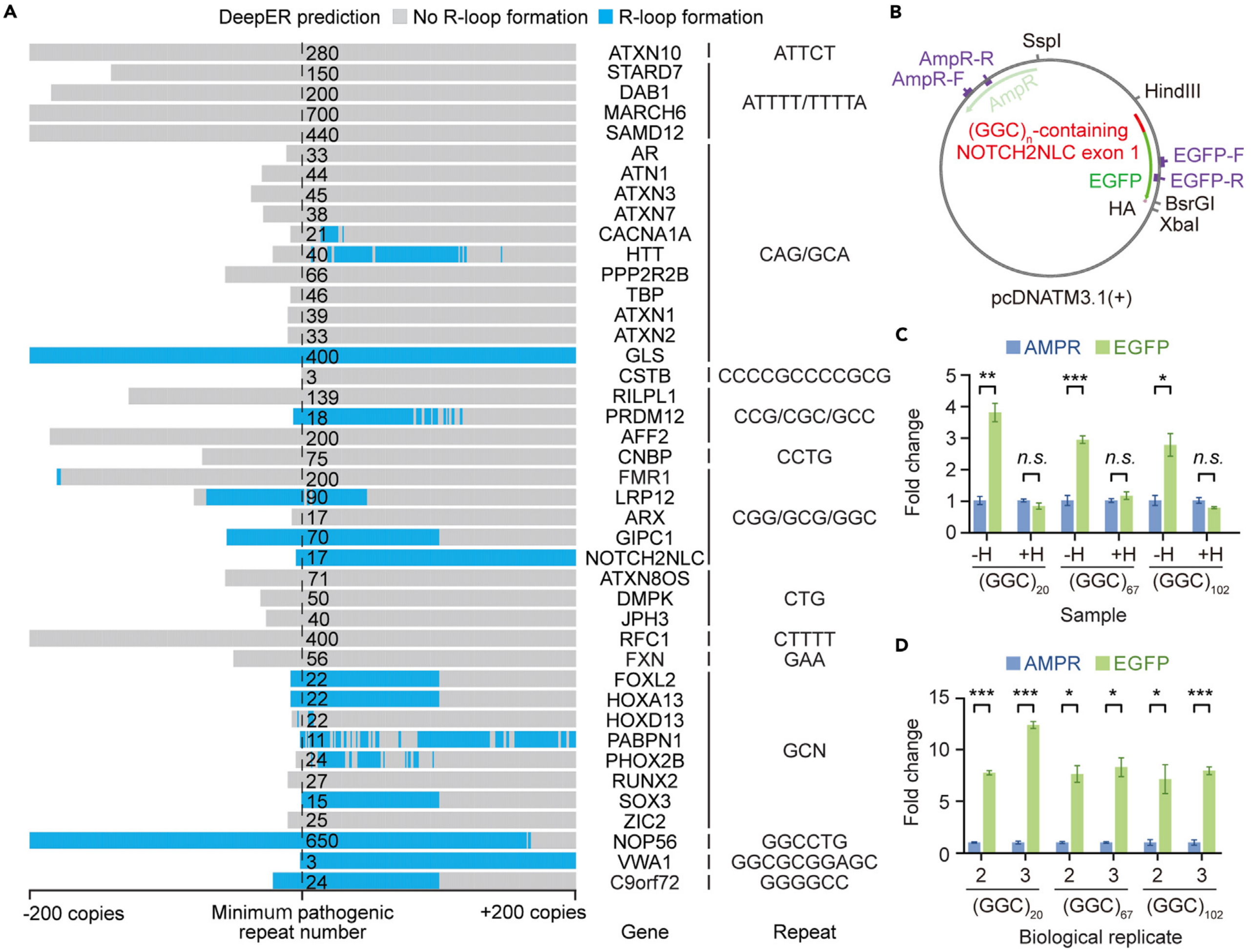


Figure 4. DeepER predicts the link of R-loop formation with some repeat expansion diseases

(A) Shown are predicted probabilities of R-loop formation (left) for indicated gene loci (middle) with different numbers of indicated tandem repeats (right) in their native sequence contexts. The minimum pathogenic repeat numbers are labeled.

(B) Expression vector of GGC repeats (20, 67, or 102 copies) within NOTCH2NLC fused with EGFP and HA tag. SspI, HindIII, BsrGI, and XbaI restriction sites, as well as PCR primers are indicated.

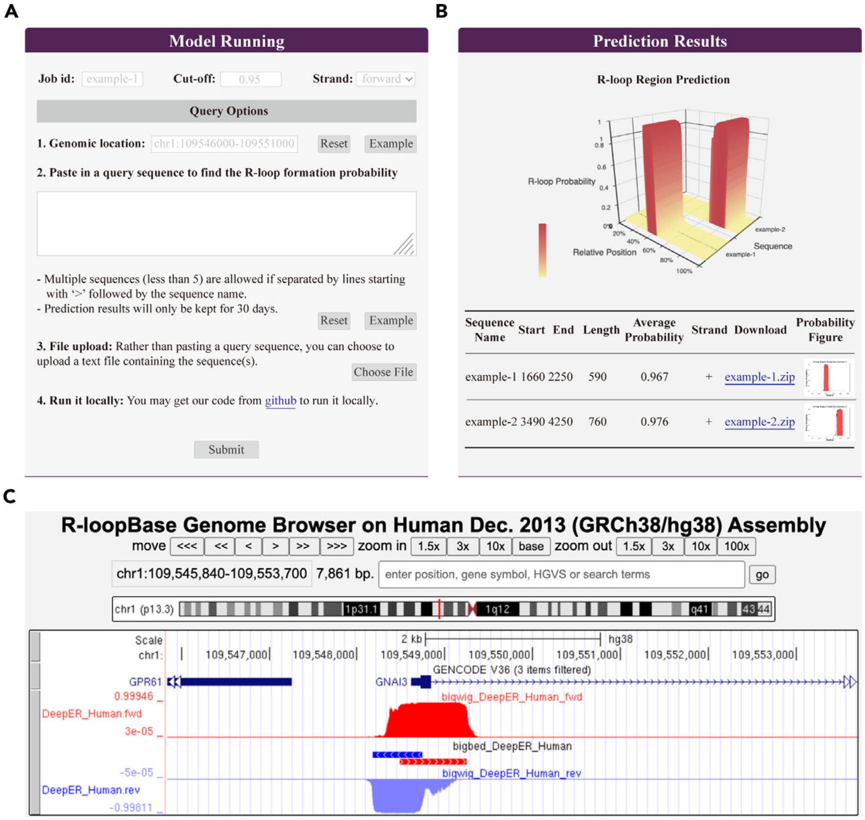
(C) DRIP-qPCR signals relative to input for AMPR and EGFR gene loci are measured and normalized to those for AMPR gene. H, RNase H1 treatment. Data are represented as mean ± SEM (n = 3). *p-value < 0.05, **p-value < 0.01 and ***p value < 0.001 determined by two-sided unpaired Student's t test. *n.s.*, not significant.

(D) DRIP-qPCR results for two additional biological replicates (2 and 3). Data are represented as mean ± SEM (n = 3). *p-value < 0.05, ***p-value < 0.001 determined by two-sided unpaired Student's t test.

pathogenic repeat number, whichever is more. These repeat sequences, along with their surrounding sequences, were then subjected to DeepER for R-loop prediction. DeepER successfully predicted R-loop formation at CAG (e.g., *HTT* gene), CGG (e.g., *FMR1* gene), GGGGCC (e.g., *C9orf72* gene) repeats that have demonstrated R-loop formation during *in vitro* transcription or at endogenous gene loci[6,34,37] (Figure 4A). This finding suggests that DeepER could be used to investigate the potential link between R-loop formation and some repeat expansion diseases.

We subsequently revealed that certain types of repeats exhibited a strong association with R-loop formation (Figure 4A). Consistent with the data presented in Figure 3F, GGC/CGG/GCG, GCN, and GGGGCC tandem repeats generally displayed a pronounced propensity for R-loop formation. Additionally, GGCGCGGAGC were observed strongly associated with R-loop formation. However, AT-rich tandem repeats, CAG/GCA, and CTG repeats were less likely associated with R-loop formation. It is noteworthy that even though different genes harbored the same copy number of the same tandem repeats, their probabilities of R-loop formation were not always the same, suggesting the influence of native sequence contexts on R-loop formation.

DeepER predicted strong R-loop formation at tandem repeats in *GIPC1*, *NOTCH2NLC*, *VWA1* genes and others. However, there is currently no experimental evidence available regarding the formation of R-loops at these genes. As a proof of concept, we constructed



Model Running

Job id: example-1 Cut-off: 0.95 Strand: forward

Query Options

1. Genomic location: chr1:109546000-109551000

2. Paste in a query sequence to find the R-loop formation probability

- Multiple sequences (less than 5) are allowed if separated by lines starting with '>' followed by the sequence name.
- Prediction results will only be kept for 30 days.

3. File upload: Rather than pasting a query sequence, you can choose to upload a text file containing the sequence(s).

4. Run it locally: You may get our code from github to run it locally.

Prediction Results

R-loop Region Prediction

Sequence Name	Start	End	Length	Average Probability	Strand	Download	Probability Figure
example-1	1660	2250	590	0.967	+	example-1.zip	
example-2	3490	4250	760	0.976	+	example-2.zip	

R-loopBase Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

chr1:109,545,840-109,553,700 7,861 bp. enter position, gene symbol, HGVS or search terms

Figure 5. DeepER web server

(A) Different options (from 1 to 4) for running DeepER.

(B) An example showing the prediction results of DeepER. Top, an interactive plot showing the distribution of R-loop formation probability. Bottom, the meta information for predicted R-loop regions.

(C) Screenshot of R-loopBase genome browser showing the pre-predicted genome-wide R-loops by DeepER.

expression vectors carrying *NOTCH2NLC* exon 1 with 20, 67 and 102 copies of GGC repeats that were fused with EGFP and HA tag (Figure 4B),[41] and verified the formation of RNase H1-sensitive R-loop structures on GGC repeats-containing restriction fragments via DRIP-qPCR (Figure 4C) in different biological replicates (Figure 4D). Altogether, our DeepER predicts the potential link of R-loop formation with some repeat expansion diseases, shedding new light on understanding the pathomechanisms.

Development of DeepER web server for customized R-loop prediction

To facilitate the utilization of DeepER for customized R-loop prediction, we developed a DeepER web server (https://rloopbase.nju.edu.cn/deepr/tool/model) as an integral component of our R-loopBase application.[21] The DeepER web server can be run in four different modes, catering to the diverse requirements of users (Figure 5A). Firstly, users have the option to provide genomic coordinates. DeepER will retrieve the corresponding genomic sequence on behalf of the users to perform R-loop prediction. Secondly, users can directly paste their query sequence(s) into the query box. The web server accepts multiple sequences as long as they are provided in the standard FASTA format. Thirdly, rather than pasting sequences into the query box, users can choose to upload a FASTA file containing one or more query sequences for R-loop prediction. Lastly, for advanced users, DeepER can be downloaded as standalone software, enabling the prediction of R-loops from sequences locally. It does not necessitate any specialized hardware environment, apart from a recommended memory size of at least 4 GB to prevent overflow. An NVIDIA GPU (graphics processing unit) is also recommended for accelerated prediction speed.

Several user-friendly options are available (Figure 5A), including the ability to accept sequences of arbitrary length as input, define a custom cutoff for the average base-level probability of R-loop formation, select the forward or reverse sequence for prediction, and input a job id to facilitate result tracking. By providing these versatile input options and customizable parameters, we aim to enhance the user experience and accommodate different preferences when utilizing DeepER for R-loop prediction.

Once a job is submitted, DeepER will return prediction results on a new page, which will be retained for a period of 30 days (Figure 5B). These results can be visualized through an interactive 3D plot displaying base-level probabilities of R-loop formation (Figure 5B, top). The meta-information of each predicted R-loop, including the start position, end position, length, average probability, and strand information, is summarized in a table (Figure 5B, bottom). Furthermore, links to position-level probability scores and editable figures are available for download, catering to customized downstream analysis and presentation needs.

Furthermore, we have generated whole-genome annotation of R-loops with DeepER. These pre-prepared results have been seamlessly integrated into the R-loopBase genome browser as an independent track (Figure 5C). This integration enables users to conveniently visualize the predicted R-loops alongside other genomic data available in R-loopBase database, and facilitate valuable insights into the potential functional implications of these structures.

DISCUSSION

Broad discrepancies exist among experimental approaches for genome-wide R-loop mapping,[1,18,20,42] and it is sometimes impractical to experimentally detect R-loops. Consequently, the need for an R-loop prediction tool arises. Deep learning models have recently shown remarkable success in the genomics field.[43] The first deep learning-based tool in R-loop field, deepRloopPre, was developed recently for predicting R-loops in plants.[31] However, the deepRloopPre model was trained solely based on ssDRIP-seq data. Since a substantial portion of ssDRIP-seq-mapped R-loops lacked support from other technologies, and there is no consensus on the best method,[18] the reliability of deepRloopPre-predicted R-loops remains uncertain. In this study, we trained the DeepER model using R-ChIP-mapped R-loops that were also supported by other technologies as positive data, while using regions undetected by any technologies as negative data. The R-loop regions predicted by our DeepER were in good agreement with those identified by other experimental or computational methods.

Formation of R-loops, characterized by specific sequence features, is conserved across species.[9] Therefore, QmRLFS-finder, initially applied in the human genome, is utilized for predicting R-loop forming sequences in other species.[26] We believe that it is feasible to employ our DeepER for R-loop predictions in other species as well.

DeepER deepens our understanding of sequence characteristics linked to R-loop formation. Although G-rich and A-rich sequences have previously been reported as associated with R-loop formation, DeepER analysis suggests that their influences are contingent on their positions. Furthermore, DeepER predicted a substantial association between specific tandem repeats and R-loop formation, suggesting that R-loop formation may serve as the underlying mechanism for certain repeat expansion diseases.

Limitations of the study

While DeepER demonstrated notable performance, there is still room for improvement. In addition to sequence features, R-loops are associated with nucleosome-free regions,[44] the presence of complementary RNA molecules and other factors. The performance of DeepER will probably get improved if trained based on both DNA sequences and epigenomics data, e.g., chromatin accessibility, DNA modifications and histone modifications. Although DeepER predicts the potential link between R-loop formation and certain repeat expansion diseases, further efforts are still needed to elucidate the underlying mechanisms. Theoretically, DeepER can be used for R-loop prediction in any species. However, experimentally determined R-loops in species other than human are generally scarce or profiled via specific R-loop mapping technologies that may suffer from a high rate of false positives. Therefore, we were unable to evaluate the performance of our DeepER in these species.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Cell lines and plasmids
- METHOD DETAILS
  - Data preparation for DeepER model
  - DeepER architecture
  - DeepER training
  - DeepER hyperparameters
  - Evaluation of DeepER
  - DeepER prediction
  - Comparison of DeepER with R-loop tracker and deepRloopPre
  - Deep learning models based on feedforward neural networks
  - Deep learning model based on U-Net
  - R-loop characterization
  - Feature importance analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2024.110584.

ACKNOWLEDGMENTS

We acknowledge all members of J.-Y.C. lab for critical reading of this manuscript. This work was financially supported by fundings from National Natural Science Foundation of China (32170653 and 22293051 to J.-Y.C.; 62376215 to L.H.), Ministry of Science and Technology of China (2021YFF1201500 to J.-Y.C.), the Fundamental Research Funds for the Central Universities (2024300417 to J.-Y.C.), Jiangsu Specially-Appointed Professor Program to J.-Y.C., and the AI & AI for Science Project of Nanjing University to J.-Y.C.

AUTHOR CONTRIBUTIONS

Conceptualization, J.-Y.C.; methodology, Z.X., H.Y., Y.Z., L.H., L.G., L.X., J.W., E.L., and J.J.; software, Z.X. and H.Y.; validation, J.H., X.Zhang., and L.C.; formal analysis, J.H., Y.Z., and J.Y.; investigation, J.H., Y.Z., and Z.X.; resources, Y.P. and Q.L.; writing – original draft, J.-Y.C. and Y.Z.; writing – review & editing, J.-Y.C., J.H., and Z.X.; visualization, J.H., H.Y., X.Zhong., and R.L.; supervision, J.-Y.C.; project administration, J.-Y.C.; funding acquisition, J.-Y.C. and L.H.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 27, 2023
Revised: April 17, 2024
Accepted: July 22, 2024
Published: July 25, 2024

REFERENCES

1. Crossley, M.P., Bocek, M., and Cimprich, K.A. (2019). R-Loops as Cellular Regulators and Genomic Threats. Mol. Cell *73*, 398–411. https://doi.org/10.1016/j.molcel.2019.01.024.
2. Santos-Pereira, J.M., and Aguilera, A. (2015). R loops: new modulators of genome dynamics and function. Nat. Rev. Genet. *16*, 583–597. https://doi.org/10.1038/nrg3961.
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. https://doi.org/10.1038/s41580-019-0206-3.
4. Petermann, E., Lan, L., and Zou, L. (2022). Sources, resolution and physiological relevance of R-loops and RNA-DNA hybrids. Nat. Rev. Mol. Cell Biol. *23*, 521–540. https://doi.org/10.1038/s41580-022-00474-x.
5. García-Muse, T., and Aguilera, A. (2019). R Loops: From Physiological to Pathological Roles. Cell *179*, 604–618. https://doi.org/10.1016/j.cell.2019.08.055.
6. Richard, P., and Manley, J.L. (2017). R Loops and Links to Human Disease. J. Mol. Biol. *429*, 3168–3180. https://doi.org/10.1016/j.jmb.2016.08.031.
7. 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. https://doi.org/10.1016/j.molcel.2012.01.017.
8. Nadel, J., Athanasiadou, R., Lemetre, C., Wijetunga, N.A., Ó Broin, P., Sato, H., Zhang, Z., Jeddeloh, J., Montagna, C., Golden, A., et al. (2015). RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. Epigenet. Chromatin *8*, 46. https://doi.org/10.1186/s13072-015-0040-6.
9. 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. https://doi.org/10.1016/j.molcel.2016.05.032.
10. Dumelie, J.G., and Jaffrey, S.R. (2017). Defining the location of promoter-associated R-loops at near-nucleotide resolution using bisDRIP-seq. Elife *6*, e28306. https://doi.org/10.7554/eLife.28306.
11. Xu, W., Xu, H., Li, K., Fan, Y., Liu, Y., Yang, X., and Sun, Q. (2017). The R-loop is a common chromatin feature of the Arabidopsis genome. Nat. Plants *3*, 704–714. https://doi.org/10.1038/s41477-017-0004-x.
12. Crossley, M.P., Bocek, M.J., Hamperl, S., Swigut, T., and Cimprich, K.A. (2020). qDRIP: a method to quantitatively assess RNA-DNA hybrid formation genome-wide. Nucleic Acids Res. *48*, e84. https://doi.org/10.1093/nar/gkaa500.
13. Chen, J.Y., Zhang, X., Fu, X.D., and Chen, L. (2019). R-ChIP for genome-wide mapping of R-loops by using catalytically inactive RNASEH1. Nat. Protoc. *14*, 1661–1685. https://doi.org/10.1038/s41596-019-0154-6.
14. Tan-Wong, S.M., Dhir, S., and Proudfoot, N.J. (2019). R-Loops Promote Antisense Transcription across the Mammalian Genome. Mol. Cell *76*, 600–616.e6. https://doi.org/10.1016/j.molcel.2019.10.002.
15. 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. https://doi.org/10.1016/j.celrep.2019.09.052.
16. Wulfridge, P., and Sarma, K. (2021). A nuclease- and bisulfite-based strategy captures strand-specific R-loops genome-wide. Elife *10*, e65146. https://doi.org/10.7554/eLife.65146.
17. Wang, K., Wang, H., Li, C., Yin, Z., Xiao, R., Li, Q., Xiang, Y., Wang, W., Huang, J., Chen, L., et al. (2021). Genomic profiling of native R loops with a DNA-RNA hybrid recognition sensor. Sci. Adv. *7*, eabe3516. https://doi.org/10.1126/sciadv.abe3516.
18. Chen, J.-Y., Lim, D.-H., Chen, L., Zhou, Y., Zhang, F., Shao, C., Zhang, X., Li, H., Wang, D., Zhang, D.-E., and Fu, X.-D. (2022). Systematic Evaluation of Different R-Loop Mapping Methods: Achieving Consensus, Resolving Discrepancies and Uncovering Distinct Types of RNA: DNA Hybrids. Preprint at bioRxiv. https://doi.org/10.1101/2022.02.18.480986.
19. Castillo-Guzman, D., and Chédin, F. (2021). Defining R-loop classes and their contributions to genome instability. DNA Repair *106*, 103182. https://doi.org/10.1016/j.dnarep.2021.103182.
20. 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. https://doi.org/10.15252/embj.2020106394.

21. Lin, R., Zhong, X., Zhou, Y., Geng, H., Hu, Q., Huang, Z., Hu, J., Fu, X.D., Chen, L., and Chen, J.Y. (2022). R-loopBase: a knowledgebase for genome-wide R-loop formation and regulation. Nucleic Acids Res. *50*, D303–D315. https://doi.org/10.1093/nar/gkab1103.
22. Miller, H.E., Montemayor, D., Abdul, J., Vines, A., Levy, S.A., Hartono, S.R., Sharma, K., Frost, B., Chédin, F., and Bishop, A.J.R. (2022). Quality-controlled R-loop meta-analysis reveals the characteristics of R-loop consensus regions. Nucleic Acids Res. *50*, 7260–7286. https://doi.org/10.1093/nar/gkac537.
23. Roberts, R.W., and Crothers, D.M. (1992). Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition. Science *258*, 1463–1466. https://doi.org/10.1126/science.1279808.
24. 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. https://doi.org/10.1016/j.molcel.2017.10.008.
25. Phoenix, P., Raymond, M.A., Massé, E., and Drolet, M. (1997). Roles of DNA topoisomerases in the regulation of R-loop formation *in vitro*. J. Biol. Chem. *272*, 1473–1479. https://doi.org/10.1074/jbc.272.3.1473.
26. 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. https://doi.org/10.1093/nar/gkw1054.
27. Stolz, R., Sulthana, S., Hartono, S.R., Malig, M., Benham, C.J., and Chedin, F. (2019). Interplay between DNA sequence and negative superhelicity drives R-loop structures. Proc. Natl. Acad. Sci. USA *116*, 6260–6269. https://doi.org/10.1073/pnas.1819476116.
28. Huppert, J.L. (2008). Thermodynamic prediction of RNA-DNA duplex-forming regions in the human genome. Mol. Biosyst. *4*, 686–691. https://doi.org/10.1039/b800354h.
29. Brázda, V., Havlik, J., Kolomaznik, J., Trenz, O., and Stastny, J. (2021). R-Loop Tracker: Web Access-Based Tool for R-Loop Detection and Analysis in Genomic DNA Sequences. Int. J. Mol. Sci. *22*, 12857. https://doi.org/10.3390/ijms222312857.
30. Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. Nat. Genet. *51*, 12–18. https://doi.org/10.1038/s41588-018-0295-5.
31. Li, K., Wu, Z., Zhou, J., Xu, W., Li, L., Liu, C., Li, W., Zhang, C., and Sun, Q. (2023). R-loopAtlas: An integrated R-loop resource from 254 plant species sustained by a deep-learning-based tool. Mol. Plant *16*, 493–496. https://doi.org/10.1016/j.molp.2022.12.012.
32. Malik, I., Kelley, C.P., Wang, E.T., and Todd, P.K. (2021). Molecular mechanisms underlying nucleotide repeat expansion disorders. Nat. Rev. Mol. Cell Biol. *22*, 589–607. https://doi.org/10.1038/s41580-021-00382-6.
33. Sakamoto, N., Chastain, P.D., Parniewski, P., Ohshima, K., Pandolfo, M., Griffith, J.D., and Wells, R.D. (1999). Sticky DNA: self-association properties of long GAA.TTC repeats in R.R.Y triplex structures from Friedreich's ataxia. Mol. Cell *3*, 465–475. https://doi.org/10.1016/s1097-2765(00)80474-8.
34. Haeusler, A.R., Donnelly, C.J., Periz, G., Simko, E.A.J., Shaw, P.G., Kim, M.S., Maragakis, N.J., Troncoso, J.C., Pandey, A., Sattler, R., et al. (2014). C9orf72 nucleotide repeat structures initiate molecular cascades of disease. Nature *507*, 195–200. https://doi.org/10.1038/nature13124.
35. Groh, M., Lufino, M.M.P., Wade-Martins, R., and Gromak, N. (2014). R-loops associated with triplet repeat expansions promote gene silencing in Friedreich ataxia and fragile X syndrome. PLoS Genet. *10*, e1004318. https://doi.org/10.1371/journal.pgen.1004318.
36. Farg, M.A., Konopka, A., Soo, K.Y., Ito, D., and Atkin, J.D. (2017). The DNA damage response (DDR) is induced by the C9orf72 repeat expansion in amyotrophic lateral sclerosis. Hum. Mol. Genet. *26*, 2882–2896. https://doi.org/10.1093/hmg/ddx170.
37. Loomis, E.W., Sanz, L.A., Chédin, F., and Hagerman, P.J. (2014). Transcription-associated R-loop formation across the human FMR1 CGG-repeat region. PLoS Genet. *10*, e1004294. https://doi.org/10.1371/journal.pgen.1004294.
38. Reddy, K., Tam, M., Bowater, R.P., Barber, M., Tomlinson, M., Nichol Edamura, K., Wang, Y.H., and Pearson, C.E. (2011). Determinants of R-loop formation at convergent bidirectionally transcribed trinucleotide repeats. Nucleic Acids Res. *39*, 1749–1762. https://doi.org/10.1093/nar/gkq935.
39. Miller, H.E., Montemayor, D., Li, J., Levy, S.A., Pawar, R., Hartono, S., Sharma, K., Frost, B., Chedin, F., and Bishop, A.J.R. (2023). Exploration and analysis of R-loop mapping data with RLBase. Nucleic Acids Res. *51*, D1129–D1137. https://doi.org/10.1093/nar/gkac732.
40. Duquette, M.L., Handa, P., Vincent, J.A., Taylor, A.F., and Maizels, N. (2004). Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA. Genes Dev. *18*, 1618–1629. https://doi.org/10.1101/gad.1200804.
41. Liu, Q., Zhang, K., Kang, Y., Li, Y., Deng, P., Li, Y., Tian, Y., Sun, Q., Tang, Y., Xu, K., et al. (2022). Expression of expanded GGC repeats within *NOTCH2NLC* causes behavioral deficits and neurodegeneration in a mouse model of neuronal intranuclear inclusion disease. Sci. Adv. *8*, eadd6391. https://doi.org/10.1126/sciadv.add6391.
42. Vanoosthuyse, V. (2018). Strengths and Weaknesses of the Current Strategies to Map and Characterize R-Loops. Noncoding RNA *4*, 9. https://doi.org/10.3390/ncrna4020009.
43. Eraslan, G., Avsec, Ž., Gagneur, J., and Theis, F.J. (2019). Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet. *20*, 389–403. https://doi.org/10.1038/s41576-019-0122-6.
44. Boque-Sastre, R., Soler, M., Oliveira-Mateos, C., Portela, A., Moutinho, C., Sayols, S., Villanueva, A., Esteller, M., and Guil, S. (2015). Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. Proc. Natl. Acad. Sci. USA *112*, 5785–5790. https://doi.org/10.1073/pnas.1421197112.
45. Zhang, K., Zhang, X., Cai, Z., Zhou, J., Cao, R., Zhao, Y., Chen, Z., Wang, D., Ruan, W., Zhao, Q., et al. (2018). A novel class of microRNA-recognition elements that function only within open reading frames. Nat. Struct. Mol. Biol. *25*, 1019–1027. https://doi.org/10.1038/s41594-018-0136-3.
46. ENCODE Project Consortium, Moore, J.E., Purcaro, M.J., Pratt, H.E., Epstein, C.B., Shoresh, N., Adrian, J., Kawli, T., Davis, C.A., Dobin, A., et al. (2020). Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature *583*, 699–710. https://doi.org/10.1038/s41586-020-2493-4.
47. 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. https://doi.org/10.1016/j.jmb.2020.02.014.
48. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359. https://doi.org/10.1038/nmeth.1923.
49. Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137. https://doi.org/10.1186/gb-2008-9-9-r137.
50. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.
51. Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME Suite. Nucleic Acids Res. *43*, W39–W49. https://doi.org/10.1093/nar/gkv416.
52. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.
53. Yu, J., Deng, J., and Wang, Z. (2022). Oculopharyngodistal myopathy. Curr. Opin. Neurol. *35*, 637–644. https://doi.org/10.1097/wco.0000000000001089.
54. Khristich, A.N., and Mirkin, S.M. (2020). On the wrong DNA track: Molecular mechanisms of repeat-mediated genome instability. J. Biol. Chem. *295*, 4134–4170. https://doi.org/10.1074/jbc.REV119.007678.
55. Stoyas, C.A., and La Spada, A.R. (2018). Chapter 11 - The CAG–polyglutamine repeat diseases: a clinical, molecular, genetic, and pathophysiologic nosology. In Handbook of Clinical Neurology, D.H. Geschwind, H.L. Paulson, and C. Klein, eds. (Elsevier), pp. 143–170. https://doi.org/10.1016/B978-0-444-63233-3.00011-7.
56. O'Hearn, E., Holmes, S.E., and Margolis, R.L. (2012). Chapter 34 - Spinocerebellar ataxia type 12. In Handbook of Clinical Neurology, S.H. Subramony and A. Dürr, eds. (Elsevier), pp. 535–547. https://doi.org/10.1016/B978-0-444-51892-7.00034-6.
57. Kurosaki, T., and Ashizawa, T. (2022). The genetic and molecular features of the intronic pentanucleotide repeat expansion in spinocerebellar ataxia type 10. Front. Genet. *13*, 936869. https://doi.org/10.3389/fgene.2022.936869.
58. Guo, P., and Lam, S.L. (2016). Unusual structures of CCTG repeats and their participation in repeat expansion. Biomol. Concepts *7*, 331–340. https://doi.org/10.1515/bmc-2016-0024.

59. Gecz, J. (2000). The FMR2 gene, FRAXE and non-specific X-linked mental retardation: clinical and molecular aspects. Ann. Hum. Genet. *64*, 95–106. https://doi.org/10.1046/j.1469-1809.2000.6420095.x.

60. Cuccurullo, C., Striano, P., and Coppola, A. (2023). Familial Adult Myoclonus Epilepsy: A Non-Coding Repeat Expansion Disorder of Cerebellar-Thalamic-Cortical Loop. Cells *12*, 1617. https://doi.org/10.3390/cells12121617.

61. Verkerk, A.J., Pieretti, M., Sutcliffe, J.S., Fu, Y.-H., Kuhl, D.P., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F.P., et al. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell *65*, 905–914. https://doi.org/10.1016/0092-8674(91)90397-H.

62. van Kuilenburg, A.B.P., Tarailo-Graovac, M., Richmond, P.A., Drögemöller, B.I., Pouladi, M.A., Leen, R., Brand-Arzamendi, K., Dobritzsch, D., Dolzhenko, E., Eberle, M.A., et al. (2019). Glutaminase Deficiency Caused by Short Tandem Repeat Expansion in GLS. N. Engl. J. Med. *380*, 1433–1441. https://doi.org/10.1056/NEJMoa1806627.

63. Maureen, G.P., Jong, W.Y., Daniel, L.B., Fang, D., Richard, A.H., James, D.F., and Jeffrey, L.N. (2009). A Triplet Repeat Expansion Genetic Mouse Model of Infantile Spasms Syndrome, Arx(GCG)10+7, with Interneuronopathy, Spasms in Infancy, Persistent Seizures, and Adult Cognitive and Behavioral Impairment. J. Neurosci. *29*, 8752–8763. https://doi.org/10.1523/JNEUROSCI.0915-09.2009.

64. Wu, Y.-R., Chen, I.C., Soong, B.-W., Kao, S.-H., Lee, G.-C., Huang, S.-Y., Fung, H.-C., Lee-Chen, G.-J., and Chen, C.-M. (2009). SCA8 repeat expansion: large CTA/CTG repeat alleles in neurological disorders and functional implications. Hum. Genet. *125*, 437–444. https://doi.org/10.1007/s00439-009-0641-x.

65. Pagnamenta, A.T., Kaiyrzhanov, R., Zou, Y., Da'as, S.I., Maroofian, R., Donkervoort, S., Dominik, N., Lauffer, M., Ferla, M.P., Orioli, A., et al. (2021). An ancestral 10-bp repeat expansion in VWA1 causes recessive hereditary motor neuropathy. Brain *144*, 584–600. https://doi.org/10.1093/brain/awaa420.

66. Cortese, A., Simone, R., Sullivan, R., Vandrovcova, J., Tariq, H., Yau, W.Y., Humphrey, J., Jaunmuktane, Z., Sivakumar, P., Polke, J., et al. (2019). Biallelic expansion of an intronic repeat in RFC1 is a common cause of late-onset ataxia. Nat. Genet. *51*, 649–658. https://doi.org/10.1038/s41588-019-0372-4.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Mouse monoclonal S9.6 antibody	Kerafast	Cat# ENH001; RRID: AB_2687463
Chemicals, peptides, and recombinant proteins		
Proteinase K	Sigma-Aldrich	Cat# 124568
Glycogen	Thermo Fisher	Cat# R0561
HindIII - HF	NEB	Cat# R3104
BsrGI - HF	NEB	Cat# R3575
XbaI	NEB	Cat# R0145
SspI - HF	NEB	Cat# R3132
RNase H	Thermo Fisher	Cat# EN0201
RiboLock RNase inhibitor	Thermo Fisher	Cat# EO0382
Experimental models: Cell lines		
Human: HEK293T	Laboratory of Yu Zhou at Wuhan University	Zhang et al.[45]
Oligonucleotides		
Primers for DRIP-qPCR experiment	This Paper	See method details
Deposited data		
R-ChIP data	Chen et al.[24]	GEO: GSE97072
K562 RNA-seq data	Moore et al.[46]	ENCODE: ENCFF671NWM
HeLa-S3 RNA-seq data	Moore et al.[46]	ENCODE: ENCFF625ZJI
HepG2 RNA-seq data	Moore et al.[46]	ENCODE: ENCFF281BBM
A549 RNA-seq data	Moore et al.[46]	ENCODE: ENCFF739RNG
DRIP-qPCR data	This Paper	Mendeley Data: https://doi.org/10.17632/4z4z2kxng8.1
Software and algorithms		
Bedtools	Quinlan et al.[47]	https://bedtools.readthedocs.io/en/latest/
Bowtie2	Langmead et al.[48]	https://bowtie-bio.sourceforge.net/bowtie2/index.shtml
DeepER	This Paper	Zenodo: https://doi.org/10.5281/zenodo.12596858
MACS2	Zhang et al.[49]	https://github.com/macs3-project/MACS
featureCounts	Liao et al.[50]	http://subread.sourceforge.net
R	N/A	https://www.r-project.org/
Python	N/A	https://www.python.org/
Pytorch	N/A	https://pytorch.org/
MEME	Bailey et al.[51]	https://meme-suite.org/meme/
R-loopBase	Lin et al.[21]	https://rloopbase.nju.edu.cn/
R-loopDB	Jenjaroenpun et al.[26]	http://r-loop.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Jia-Yu Chen (jiayuchen@nju.edu.cn).

Materials availability

The plasmids and cell lines generated in this study are available upon request.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. Original qPCR results have been deposited at Mendeley Data and are publicly available as of the date of publication. The DOI is listed in the key resources table.
- All original code has been deposited at Zenodo and are publicly available as of the data of publication. DOI is listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Cell lines and plasmids

HEK293T cells, previously purchased from and authenticated by the Cell Bank of the Chinese Academy of Sciences (Shanghai, China), were obtained from the laboratory of Yu Zhou at Wuhan University[45] and cultured in DMEM supplemented with 10% FBS and penicillin/streptomycin (100 U/ml) at 37°C with 5% CO_2. Cells were determined to be free from mycoplasma contamination. Expression vector carrying *NOTCH2NLC* exon1 with different numbers of GGC repeats were prepared and transfected into HEK293T cells as described before.[41]

METHOD DETAILS

Data preparation for DeepER model

R-ChIP data from HEK293 and K562 cells were downloaded from the GEO database (GEO: GSE97072).[24] All data were aligned to the human genome (hg38) using Bowtie2,[48] retaining only uniquely-mapped reads. MACS2 was employed with stringent criteria (fold change ≥ 5 and q-value ≤ 0.001) to identify R-loop peaks.[49] Peaks exhibiting a 50% reciprocal overlap between the two cell lines were merged as conservative R-loop regions. R-loop regions were considered only if they were supported by $\geq$ 1 R-loop dataset generated by other R-loop mapping technologies achieved in R-loopBase.[21]

Input data for model training were the prepared as follows. We randomly selected 5-kb-long intervals containing the aforementioned R-loop peaks and surrounding R-loop-negative regions across the human genome. We then introduced a data augmentation step, before which the testing set was separated in advance to ensure there would be no data leakage. Ten intervals for each R-loop region were selected to enhance the model's robustness in handling R-loops at various positions relative to the 5-kb segment. Base positions within R-loop regions were assigned as 1 and other positions as 0. An approximately equivalent number of intervals were randomly selected from the entire genome to serve as negative intervals. These negative intervals had no overlap with the positive intervals, gap regions, or peaks detected by any other datasets in R-loopBase.[21] All bases for negative intervals were labeled as 0. All intervals were allocated into training, validation and testing datasets at a ratio of 7:2:1. Sequences were then extracted and encoded using one-hot encoding for model training.

DeepER architecture

The DeepER architecture includes a standard BiLSTM layer, two BiLSTM layers with residual connections, and a fully connected layer activated by sigmoid function which normalizes the output to the range of [0, 1]. The probability of R-loop formation, *Y*, is computed as a function of input sequence, *X*:

$$Y = DeepER(X)$$

To elaborate further, the computational steps of the DeepER model can be expressed in the following pseudocode:

```
$X_{OneHot} = OneHot(X)$
$X_0 = BiLSM_0(X_{OneHot})$
for i =1 to 2:
    $X_i = resBiLSM_i(X_{i-1})$
$Y = Sigmoid(FC(X_2))$
return Y
```

OneHot is a function to convert the input sequence into its one-hot encoding representation. *BiLSTM*$_0$ is a standard BiLSTM layer. *ResBiLSTM*$_i$ is the *i*-th layer of BiLSTM with a residual connection. *FC* is a fully connected layer. *Sigmoid* is an activation function typically used for binary classification tasks that maps input values to a range of [0, 1].

Specifically, BiLSTM denotes a bidirectional LSTM layer that processes input sequences in both forward and backward directions. Forward LSTM equations are defined as follows:

$$f_i = \sigma(W_f \cdot [h_{i-1}, x_i] + b_f)$$

$$i_i = \sigma(W_i \cdot [h_{i-1}, x_i] + b_i)$$

$$\tilde{C}_i = tanh(Wc \cdot [h_{i-1}, x_i] + b_C)$$

$$C_i = f_i \odot C_{i-1} + i_i \odot \tilde{C}_i$$

$$o_i = \sigma(W_o \cdot [h_{i-1}, x_i] + b_o)$$

$$h_i = o_i \odot \tanh(C_i)$$

Backward LSTM equations are defined as follows:

$$f'_j = \sigma\left(W'_f \cdot \left[h'_{j+1}, x_j\right] + b'_f\right)$$

$$i'_j = \sigma\left(W'_i \cdot \left[h'_{j+1}, x_j\right] + b'_i\right)$$

$$\tilde{C}'_j = tanh\left(W'c \cdot \left[h'_{j+1}, x_j\right] + b'_C\right)$$

$$C'_j = f'_j \odot C'_{j+1} + i'_j \odot \tilde{C}'_j$$

$$o'_j = \sigma\left(W'_o \cdot \left[h'_{j+1}, x_j\right] + b'_o\right)$$

$$h'_j = o'_j \odot \tanh\left(C'_j\right)$$

The BiLSTM output at each time step can be obtained by concatenating the forward and backward hidden states:

$$y_k = \left[o_k, o'_k\right]$$

In the above equations, σ denotes the sigmoid activation function, $\odot$ denotes element-wise multiplication (Hadamard product), and W and b represent the weight matrix and bias vector, respectively. h_i and h'_j represent the forward and backward hidden states at time step i and j. x_i and x_j represent the input feature at time step i and j. f_i, i_i, o_i, and $\tilde{c}_i$ represent the forget gate, input gate, output gate, and candidate memory cell at time step i, respectively. f'_j, i'_j, o'_j, and $\tilde{c}'_j$ represent the corresponding gates and cell for the backward LSTM. C_i and C'_j are responsible for updating their respective cell states. y_k denotes the final output of BiLSTM at time step k.

The residual Bi-LSTM is expressed as the following:

$$resBiLSTM(x) = x + ReLU(BiLSTM(x))$$

$$ReLU(x) = max(0, x)$$

ReLU is a rectifier linear unit activation function.

DeepER training

The model uses the Weighted Symmetric Cross Entropy loss function to measure the difference between the model prediction results and the actual labels. For the binary classification problem, the cross-entropy loss function is expressed as:

$$CE\left(y_i, y_{pred,i}\right) = L\left(y_i, y_{pred,i}\right) = -\left(y_i * \log\left(y_{pred,i}\right) + (1 - y_i) * \log\left(1 - y_{pred,i}\right)\right)$$

y_i is the actual label (0 or 1) , and $y_{pred,i}$ is the predicted probability by the model. This loss function measures the error of the model by calculating the negative logarithm of the prediction probability corresponding to the actual label.

From this, the Symmetric cross-entropy loss can be defined as:

$$SCE\left(y_i, y_{pred,i}\right) = \frac{CE\left(y_i, y_{pred,i}\right) + CE\left(y_{pred,i}, y_i\right)}{2}$$

Moreover, we have assigned different weights to the SCE loss at various positions in order to mitigate the effects of excessive negative samples and varying lengths of positive examples. The negative weights are set as 1 and the positive weights are set inversely proportional to the length of each sequence. The formula is written as follow:

$$\omega_i = \begin{cases} \frac{l}{l_p} if y_i \text{ belongs to positive sample} \\ 1 \text{ if } y_i \text{ belongs to negative sample} \end{cases}$$

l_p represents the R-loop length where the y_i belongs to.

The total loss function can be written as follow:

$$L\left(y, y_{pred}\right) = \frac{1}{l}\sum_{i=1}^{l} \omega_i \cdot SCE\left(y_i, y_{pred,i}\right) = \frac{1}{l}\sum_{i=1}^{l} \omega_i \cdot \frac{CE\left(y_i, y_{pred,i}\right) + CE\left(y_{pred,i}, y_i\right)}{2}$$

In addition, we use the Adam optimization algorithm to minimize loss and update the model's parameters. Adam combines the idea of gradient descent with the ability to adapt learning rate and has fast convergence speed and good performance. Adam's optimization algorithm formula are as follows:

$$m_t = \beta_1 * m_{t-1} + (1 - \beta_1) * g$$

$$v_t = \beta_2 * v_{t-1} + (1 - \beta_2) * g * g$$

$$\widehat{m}_t = m_t / \left(1 - \beta_1^t\right)$$

$$\widehat{v}_t = v_t / \left(1 - \beta_2^t\right)$$

$$\theta_t = \theta_{t-1} - lr_t * \widehat{m}_t / \left(\sqrt{\widehat{v}_t} + \varepsilon\right)$$

m_t is the first-order moment estimation of the gradient, v_t is the second-order moment estimation of the gradient, $\widehat{m}_t$ and $\widehat{v}_t$ are correcting for deviations respectively, β_1 and β_2 are the decay rates used to control the first-order moment and second-order moment estimations, t represents the current iteration steps, lr is the learning rate, and ε is a small constant used to prevent division by zero errors.

DeepER hyperparameters

DeepER hyperparameters were tuned experientially through sequential exploration of the hyperparameter space over the validation set. The learning rate is set as 1.6×10^{-3} and decays by 10% every 5 epochs, batch size as 64, number of epochs as 100. Initial weights were initialized with orthogonal initialization. The gradient exponential decay rate and numerical stability parameter were set as default values.

Evaluation of DeepER

Bases with predicted probability values ≥ 0.95 were classified as R-loop-positive bases. Comparing these predictions against the true labels enabled us to construct a confusion matrix, from which base-level evaluation metrics were derived. A sliding window approach was adopted to define R-loop regions. Considering that the average R-loop size is a few hundred nucleotides,[40,47] we set the window size to 200 bp and step size to 10 bp. If the average probability was ≥ 0.95, the window was defined as an R-loop region. If a segment containing true labels was intersected with predicted R-loop-forming regions, it would be considered as a true positive event. Similarly, true negative, false positive, and false negative events were also counted for calculation of region-level evaluation metrics.

DeepER prediction

A pre-processing step for query sequences was implemented to allow DeepER to accept sequences of arbitrary length. Query sequences shorter than 5 kb will be automatically extended to 5 kb by padding both ends with blanks. Sequences longer than 5 kb will be split into 5-kb segments. If a sequence is not a multiple of 5 kb, the final segment along with upstream sequences will be used for R-loop prediction for the final segment. Genome-wide annotation of R-loops were generated as follows. For each chromosome, the best DeepER model was used to predict R-loop formation probabilities of both strands. R-loop regions were predicted using a sliding window approach (window size = 200 bp, step size = 10 bp and average probability ≥ 0.95) as described above. All overlapping R-loop regions on the same strand were then merged.

Comparison of DeepER with R-loop tracker and deepRloopPre

All three tools were employed to predict R-loops across the human genome using default parameters. The genome sequences were divided into 200 bp segments. A segment was deemed a true positive event if it contained both predicted R-loops and consensus R-loops deduced by

RLBase[39] or level-4 R-loops defined by R-loopBase.[21] Similarly, true negative, false positive, and false negative events were also counted for calculation of precision, recall, F1 score, accuracy, and specificity. To assess efficiency, we utilized all three tools to predict R-loop formation of a 24-Mb long random sequence on a CentOS Linux 7 (Core) with Intel(R) Xeon(R) Silver 4210R CPU @ 2.40 GHz. Each experiment was repeated 10 times, and the averaged running time and memory size were calculated. The efficiency of DeepER was additionally evaluated on 2 × NVIDIA GeForce RTX 3090.

Deep learning models based on feedforward neural networks

In comparison with DeepER model, we built four fully connected neural networks that can be expressed as the following pseudocode:

```
$X_0 = OneHot(X)$
For i = 1 to layer_number:
    $X_i = ReLU(W_i \times X_{i-1} + b_i)$
$Y = Sigmoid\left(X_{layer\_number}\right)$
Return Y
```

The W_i and b_i represents the weight matrix and the bias of each layer. *ReLU* is an activation function. Sigmoid is used to normalize the output to the range of [0, 1]. Step $W_i \times X_{i-1} + b_i$ is the most common form of matrix multiplication and vector addition. Model 1 consists of only input and output layers. Model 2 includes an additional hidden layer. Model 3 is similar to Model 2, but the neural number of the hidden layer is smaller than that of Model 2. Model 4 has two hidden layers. The architectures of these models are illustrated in Figure S2A.

Deep learning model based on U-Net

This model is built upon the U-Net architecture, consisting of a series of encoder blocks and decoder blocks connected through skip-connections (see Figure S2).

The U-Net model can be expressed in the following pseudocode:

```
$X_0 = OneHot(X)$
For i = 1 to 4:
    $X_i = Encoderblock_i(X_{i-1})$
$X'_0 = X_4$
For i = 1 to 3:
    $X'_i = Decoderblock_i\left(X'_{i-1}, X_{4-i}\right)$
$Y = Sigmoid(X'_3)$
Return Y
```

The *Encoderblock* contains a single convolution layer and two residual blocks, and each includes two convolution layers and a batch normalization layer. It compresses the input dimensions but preserves more features. The formula is as follow:

$$Conv(x) = W * x + b$$

$$BN(x) = \gamma \cdot \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta$$

$$Residual(x) = x + ReLU(Conv(BN(Conv(x))))$$

$$Encoderblock(x) = Residual_2(Residual_1(Conv(x)))$$

σ is the variance of x, μ is the mean value of x. ε is a small number to make sure that denominator will not devide by zero. W, b, γ, β are parameters to learn.

The *Decoderblock* contains a single transposed convolution layer and two residual blocks, and each includes two convolution layers and a batch normalization layer. Meanwhile, it expands the compressed feature representations from the Encoder, while also incorporating the saved feature maps from the corresponding Encoder blocks. The formula is as follow:

$$TransConv(x) = x * W^T + b$$

$$BN(x) = \gamma \cdot \frac{x - \mu}{\sqrt{\sigma^2 + \varepsilon}} + \beta$$

$$Residual(x) = x + ReLU(Conv(BN(Conv(x))))$$

$$Encoderblock(x) = Residual_2(Residual_1(TransConv(x + x_{enc})))$$

Similarly, sigmoid is used to activate the output. The model employs the same weighted symmetric cross entropy loss function to adjust the predicted results towards the true labels as closely as possible. The Adam optimizer is utilized to efficiently update the model parameters.

R-loop characterization

To compare DeepER with QmRLFS-finder, we collected RLFS data from R-loopDB,[26] converted genomic coordinates using LiftOver and calculated their overlaps with R-loops predicted by DeepER using BEDTools.[52] Public RNA-seq data (ENCFF671NWM for K562, ENCFF625ZJI for HeLa-S3, ENCFF281BBM for HepG2 and ENCFF739RNG for A549 cells) were downloaded from ENCODE project,[46] and used for gene expression level quantification for class I and II R-loops with featureCounts.[50] STREME (default parameter) of MEME Suite[51] were used to find R-loop motifs. G4 motif and G4 ChIP-seq data were downloaded from R-loopBase[21] for comparison with DeepER-predicted R-loops.

Feature importance analysis

We used a permutation-based method to analyze the sequence features important for DeepER-predicted R-loops. In short, one single point mutation (the reference base was mutated to one of other three bases) was introduced randomly within each R-loop region using in-house scripts, followed by DeepER predictions. Mutations were considered as R-loop disrupting mutations if the resulted R-loops after mutation introduction exhibited < 50% overlap with original R-loop regions. Percentages of mutated bases, mutation types and 3-mers were determined for R-loop-disrupting and -preserving mutations for fold enrichment calculation.

R-loop predictions for disease-related tandem repeats

Information for repeat expansion diseases were collected from reviews[53–60] and additional PubMed literature[61–66] (see Table S4). Five kilobase sequences, consisting of different numbers of disease-related tandem repeats and equal number of bases upstream and downstream of repeats, were subjected to DeepER predictions. The number of random repeats ranges from 200 copies less than the minimum pathogenic repeat number or the repeat number on the reference genome to 200 copies beyond the reported minimum pathogenic number.

DRIP-qPCR validation

For DRIP-qPCR, HEK293T cells (5×10^6) were washed in cold PBS, treated with 1 ml PBS and collected by centrifuge at 600 g for 5 min at 4°C. Cells were treated with PK buffer (100 mM NaCl, 10 mM Tris pH 8.0, 1 mM EDTA, 0.5% SDS), 6 μl Proteinase K (300 μg/ml) and 3 μl Ribolock RNase inhibitor, followed by incubation at 37°C for 5 h. DNA was extracted by phenol-chloroform-isoamyl alcohol in light phase lock tubes, precipitated with 1.5 μl glycogen, 1/10 volume sodium acetate (40 μl) and 2.5-fold volume of ethanol (1,000 μl) at -80°C for 30 min, spin at 14,000 rpm at 4°C for 15 min, washed twice with 70% ethanol and re-suspend in 50 μl Tris elution buffer (10 mM Tris-HCl pH 8.0).

DNA was digested by 3 μl HindIII (20,000 units/ml, NEB), 3 μl BsrGI (20,000 units/ml, NEB), 3 μl XbaI (20,000 units/ml, NEB) and 3 μl SspI (20,000 units/ml, NEB). For R-loop validation, fragmented DNA was pretreated with 3 μl RNase H (0297S, 10-unit total NEB) at 37°C. Digested DNA was purified by 200 μl TE buffer (10 mM Tris-Cl pH 8.0, 1 mM EDTA pH 8.0), extracted by phenol-chloroform-isoamyl alcohol, followed by ethanol precipitation.

Immunoprecipitations were performed by diluting 4 μg of fragment DNA to 150 μl 1× binding buffer (10 mM $NaPO_4$ pH 7, 140 mM NaCl, 0.05% Triton X-100, 1× PI and Ribolock) and 0.4 μg withdrawn to serve as input in qPCR. RNA-DNA hybrids were immunoprecipitated with 3 μg of S9.6 overnight at 4°C.

Magnetic beads were washed 3 times with ChIP-dilution Buffer, incubated with Blocking buffer at RT for 2 hours on the rotating platform and washed 3 times with BSA+PBS. After removing BSA+PBS, DNA/antibody complex was added to beads and incubated for 2-3 hours at 4°C. Beads was washed three times in binding buffer (+0.3 × PI, 2 μl Ribolock) and elution was performed in 150 μl elution buffer (10 mM Tris pH 8, 1 mM EDTA, 1% SDS and 6 μl Proteinase K) for 45 min at 55°C. After adding 150 μl TE buffer, DNA was extracted by phenol-chloroform-isoamyl alcohol, followed by ethanol precipitation. Eluted DRIP DNA was washed twice with ethanol, re-suspended in 50 μl H_2O and analyzed by qPCR. Primers used for DRIP-qPCR were CAATGATACCGCGAGACCCA (AmpR-F), CTTGATCGTTGGGAACCGGA (AmpR-R), AAGGACGACGGCAACTACAA (EGFP-F) and CGATGTTGTGGCGGATCTTG (EGFP-R).

QUANTIFICATION AND STATISTICAL ANALYSIS

Data were shown as mean ± SEM. Statistical analysis was performed with R. Differences of mean values and frequencies between two groups were test using two-sided unpaired Student's t-test and chi-square test, respectively, as indicated in the figure legends. P-values below 0.05 were considered significant, specifically, **p*-value < 0.05, ***p*-value < 0.01, ****p*-value < 0.001 and *n.s.* stands for no significant difference. All experiments have been independently performed for three times.

ADDITIONAL RESOURCES

DeepER web server: https://rloopbase.nju.edu.cn/deepr/tool/model.