Published online 18 November 2021 *Nucleic Acids Research, 2022, Vol. 50, Database issue* ***D303–D315***
https://doi.org/10.1093/nar/gkab1103

R-loopBase: a knowledgebase for genome-wide R-loop formation and regulation

Ruoyao Lin[1,†], Xiaoming Zhong[2,*,†], Yongli Zhou[1,†], Huichao Geng[3], Qingxi Hu[1], Zhihao Huang[3], Jun Hu[1], Xiang-Dong Fu[4], Liang Chen[3,*] and Jia-Yu Chen [1,*]

[1]State Key Laboratory of Pharmaceutical Biotechnology, School of Life Sciences, Chemistry and Biomedicine Innovation Center (ChemBIC), Nanjing University, Nanjing 210023, China, [2]Ben May Department for Cancer Research, University of Chicago, Chicago, IL 60637, USA, [3]Hubei Key Laboratory of Cell Homeostasis, RNA Institute, College of Life Sciences, Wuhan University, Wuhan 430072, China and [4]Department of Cellular and Molecular Medicine, Institute of Genomic Medicine, University of California, San Diego, La Jolla, CA 92093, USA

Received August 11, 2021; Revised September 28, 2021; Editorial Decision October 18, 2021; Accepted October 21, 2021

ABSTRACT

R-loops play versatile roles in many physiological and pathological processes, and are of great interest to scientists in multiple fields. However, controversy about their genomic localization and incomplete understanding of their regulatory network raise great challenges for R-loop research. Here, we present R-loopBase (https://rloopbase.nju.edu.cn) to tackle these pressing issues by systematic integration of genomics and literature data. First, based on 107 high-quality genome-wide R-loop mapping datasets generated by 11 different technologies, we present a reference set of human R-loop zones for high-confidence R-loop localization, and spot conservative genomic features associated with R-loop formation. Second, through literature mining and multi-omics analyses, we curate the most comprehensive list of R-loop regulatory proteins and their targeted R-loops in multiple species to date. These efforts help reveal a global regulatory network of R-loop dynamics and its potential links to the development of cancers and neurological diseases. Finally, we integrate billions of functional genomic annotations, and develop interactive interfaces to search, visualize, download and analyze R-loops and R-loop regulators in a well-annotated genomic context. R-loopBase allows all users, including those with little bioinformatics background to utilize these data for their own research. We anticipate R-loopBase will become a one-stop resource for the R-loop community.

INTRODUCTION

R-loops are three-stranded nucleic acid structures composed of an RNA:DNA hybrid and a displaced single-stranded DNA (1). Initially considered as rare by-products of transcription, R-loops are now found widely distributed across genomes in species from bacteria to human (2). Excessive R-loops are critical sources of genome instability (3,4), and underlie many human diseases (5), such as cancers (6,7), neurodegenerative disorders (8) and autoimmune diseases (9). Intriguingly, R-loops have been increasingly appreciated as key cellular regulators in many physiological processes (1,4), including DNA replication (10), homologous recombination (11), DNA damage repair (12) and transcription (13). Collectively, the functional studies of R-loops have greatly advanced both basic and translational research.

Over the past decade, the recognition of the functional importance of R-loops has accelerated the development of more than ten different genome-wide R-loop detection technologies that are based on either the anti-RNA:DNA hybrid monoclonal antibody S9.6 (14–21) or (the hybrid binding domain of) catalytically-deficient RNase H1 (21–25). Although all these technologies are able to create a genome-wide R-loop map, they are not always consistent with each other regarding R-loop sizes and locations, and other associated genomic features (1,26–28). While an early electron microscopy study suggested that individual R-loop structures are of ∼200 nt in length (29), sizes of mapped R-loops are within a much broader range from a few hundred bases to several kilobases. Promoter regions are generally considered as R-loop hotspots (21), a large fraction of R-loops mapped by some technologies are however within gene bodies (18), in the vicinity of transcription termination sites (18), or even in intergenic regions (15,19).

*To whom correspondence should be addressed. Tel: +86 18061496681; Email: jiayuchen@nju.edu.cn
Correspondence may also be addressed to Liang Chen. Email: liang_chen@whu.edu.cn
Correspondence may also be addressed to Xiaoming Zhong. Email: xiaomingzhong@uchicago.edu
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

Furthermore, many mapped R-loop regions do not seem to comply with the GC skew sequence property, the G4 formation propensity, and the topological requirement, all of which have been found tightly associated with R-loop formation (13,21,29,30). While limitations of different technologies that could largely explain these discrepancies have been extensively discussed (1,26–28), what is missing is a systematic effort to compare different technologies side by side in a well-annotated genomic context, and to integrate all available R-loop mapping data to generate a reference set of R-loops for future functional R-loop studies.

An increasing number of proteins are now thought to regulate R-loop dynamics. Certain RNA processing proteins and DNA damage-related factors may counteract R-loop formation, and some nucleases and helicases may resolve existing R-loops. However, the molecular mechanism and the global regulatory network involved are only partially understood. For example, considering their pervasive association with chromatin (31), RNA processing proteins may regulate R-loops in an RNA-independent way rather than through RNA binding activity as proposed before (4). Although helicases are in general considered to limit R-loops by resolving RNA:DNA hybrids, they may instead promote R-loop formation by resolving structured RNA to facilitate its invasion into DNA (32). R-loop regulatory proteins, named as R-loop regulators, have been systematically profiled by either S9.6 antibody or hybrid probe enrichment coupled with mass spectrometry analysis (33,34). However, due to the affinity of S9.6 to dsRNA (35), it remains unclear whether the identified proteins are truly involved in R-loop regulation. Furthermore, validated R-loop regulators are scattered in the literature and how they are functionally connected is unclear.

Here, we establish R-loopBase to tackle the above challenges. The massive amount of human genomics data enables us to perform integrative analysis. We thus generate human R-loop zones of different confidence levels and R-loop regulome with comprehensive functional annotations. We also curate and annotate R-loop regulators in mouse, yeast (*Saccharomyces cerevisiae*) and *Escherichia coli* to support R-loop research in these model organisms. User-friendly interfaces are developed to allow users, even those with little bioinformatics background, to leverage these data for their own R-loop research. We will continuously update R-loopBase in the future to better serve the R-loop community.

MATERIALS AND METHODS

Collection and analyses of genome-wide R-loop mapping data in human

We collected meta information of genome-wide R-loop mapping data from all PubMed literature with the keyword "*R-loop" OR "R-loops" OR "RNA DNA hybrid" OR "RNA DNA hybrids" OR "DNA RNA hybrid" OR "DNA RNA hybrids"* as query. In current release of R-loopBase, datasets generated from human cells under basal conditions published on and before March 31st 2021 were downloaded (Supplementary Table S1). In total, 118 datasets generated by 11 different technologies were collected (9,13,14,16,18,19,21,23–25,28,34,36–55) (Table 1 and Supplementary Figure S1), and subjected to a standardized data analysis pipeline (Supplementary Figure S2). Briefly, technical replicates if existed were merged first, and raw sequencing data were then mapped to the human genome (hg38) using Bowtie2 local alignment mode (56). Uniquely-mapped non-redundant reads were kept as useful reads and samples with >7M useful reads were considered as with sufficient read counts. To maximally leverage the sequencing data, biological replicates with <7M useful reads were merged to meet with the minimal reads count cutoff as long as they were highly correlated (*Spearman* correlation coefficient > 0.5). Finally, peak calling was done with MACS2 (57) for all useful reads (DRIP-seq, DRIVE-seq, MapR and R-loop CUT&Tag) or useful reads from Watson or Crick strand separately (DRIPc-seq, RDIP-seq, ssDRIP-seq, qDRIP-seq, R-ChIP and RR-ChIP), using *q*-value cutoff 0.01 for narrow peak (R-ChIP and R-loop CUT&Tag) and 0.05 for broad peak (DRIP-seq, DRIPc-seq, RDIP-seq, ssDRIP-seq, qDRIP-seq, DRIVE-seq, MapR and RR-ChIP). If multiple biological replicates existed, peaks with ≥50 bp overlap among ≥2 replicates were merged and taken as reproducible peaks. Only peaks with strong signal enrichment (fold change ≥ 2) and outside of ChIP-seq blacklisted regions were used for downstream analysis. Samples with <100 peaks called were discarded. Following ENCODE guidelines for ChIP-seq data analyses (58), we further calculated signal portion of tags (SPOT) and reads in blacklisted regions (RiBL) as part of quality control matrix for users' reference (Supplementary Table S2). When processing bisDRIP-seq data, rather than peak calling, we uploaded the processed bisDRIP-seq data onto the R-loopBase genome browser for visualization and comparison with other data.

Generation of human R-loop zones of different confidence levels

Considering that R-loop peaks co-detected independently by multiple technologies are more likely conservative *bona fide* R-loops, we performed integrative analysis of all mapped R-loop peaks in human cells identified by all technologies. First, stranded R-loop peaks were merged as R-loop zones on Watson or Crick strand separately. Non-stranded R-loops, if overlapped with stranded R-loop peaks, were also assigned to Watson or Crick R-loop zones accordingly. The remaining non-stranded R-loop peaks were merged as non-stranded R-loop zones. Second, the resulting R-loop zones were further partitioned into subregions of *n* different confidence levels, where *n* is the minimal number of technologies by which individual subregions were detected by. For example, R-loop zones of level 3 were those co-detected by ≥3 technologies. Small R-loop zones (<50 bp) were not included.

Collection of known R-loop regulators

To collect known proteins involved in R-loop regulation from scattered literature, we downloaded all PubMed publications using keyword "*R-loop" OR "R-loops" OR "RNA DNA hybrid" OR "RNA DNA hybrids" OR "DNA RNA hybrid" OR "DNA RNA hybrids"* as query for manual cura-

Table 1. Genome-wide R-loop mapping data in human

Technology[a]	Treatment	Samples	Datasets	References
DRIP-seq	control	B-cell (1/1[b]), CHLA10 (1/1), EWS502(1/1), HeLa (4/4), HEK293 (2/2), SHSY5Y (2/2), TC32 (1/1), Stromal (4/4), Basal-epithelial (4/4), Luminal-progenitor (4/4), Mature-luminal-epithelial (4/4), MCF-7 (1/1), NT2 (6/6), K562 (2/1), Primary-fibroblast (2/2), U2OS (8/7), U87 (2/2), Jurkat (2/0), T-cells (2/0), IMR-90 (1/0), HEK293T (1/0)	55/47	(9,18,21,28,34,36–51)
	knock down	U2OS (8/6), U87 (2/2), HeLa (4/4), HEK293 (2/2), SHSY5Y (2/2)	18/16	(38,41–43,50,51)
RDIP-seq	control	HeLa (2/2), IMR-90 (1/1), HEK293T (1/0)	4/3	(19,52)
	knock down	HeLa (2/2)	2/2	(52)
DRIPc-seq	control	K562 (2/2), HEK293 (2/2), NT2 (2/2)	6/6	(18,41,48)
	knock down	K562 (2/2), HEK293 (2/2)	4/4	(41,48)
ssDRIP-seq	control	HeLa (3/3), hVECs (2/2), hESCs (2/2), hiPSCs (2/2), hMSCs (2/2), hNSCs (2/2), hVSMCs (2/2)	15/15	(53,54)
	knock down	HeLa (3/3)	3/3	(53)
bisDRIP-seq	control	MCF-7 (13/13)	13/13	(16)
qDRIP-seq	control	HeLa (3/2)	3/2	(14)
DRIVE-seq	control	NT2 (1/1)	1/1	(21)
R-ChIP	control	HEK293T (5/5), K562 (2/2), HeLa (1/0)	8/7	(13,55)
RR-ChIP	control	HeLa (2/2)	2/2	(23)
MapR	control	HEK293 (3/3), U87T (2/2)	5/5	(25)
R-loop CUT&Tag	control	HEK293T (6/6)	6/6	(24)
Sum		-	145/132	

[a]refer to Supplementary Figure S1 for procedures of different technologies.
[b]number of datasets analyzed/high-quality datasets.

tion of R-loop regulators. A protein is considered as an R-loop regulator if it binds to, stabilizes, resolves or degrades RNA:DNA hybrids, or levels of R-loops or RNA:DNA hybrids are changed upon chemical or genetic manipulation of the protein. Validated R-loop regulators in human, mouse, yeast (*Saccharomyces cerevisiae*) and *E. coli* were collected and annotated (Supplementary Table S3), and putative R-loop regulators identified in high throughput screening studies (33,34) were also included.

Protein–protein Interaction (PPI) network and gene list enrichment analysis of R-loop regulators

GO (gene ontology) (59) and DO (disease ontology) (60) enrichment analysis of validated or putative R-loop regulators were done by clusterProfiler (61), and only top ranked GO or DO terms were shown. For validated R-loop regulators, high-confidence physical interactions (minimal required interaction score = 0.700) with experimental evidences and integrated information from databases were retrieved from the STRING database (62) and visualized by Cytoscape software (63). Proteins were organized into different clusters using the AutoAnnotate plug-in (64). GO and KEGG enrichment analyses were then performed for clusters consisted of at least three genes.

Identification of R-loops targeted by R-loop regulators

R-loops have been mapped with replicates before and after gene knockdown for ten R-loop regulators (Supplementary Table S1), allowing us to identify their targeted R-loops. To do so, broad peaks were first called by referring to the references where these data were originally reported. Differential binding analyses were then done with default settings of DiffBind package (65) except for *summits = FALSE* and *bUseSummarizeOverlaps = TRUE*. DEseq2 (66) was called by DiffBind for differential analyses. Thirty-three validated human regulators have ChIP-seq data available, and 21 with CLIP-seq available. We downloaded peak files of these data from ENCODE project (67). R-loops intersected with ChIP-seq or CLIP-seq peaks as determined by BEDTools (68) were taken as potential targets of the corresponding regulator. For the above human R-loop regulators, we turned to published literature to summarize a list of their targeted R-loops validated by DNA:RNA hybrid immunoprecipitation qPCR (Supplementary Table S4). Primer-blast was used to locate the targeted R-loop region, which would be kept only when there was only one perfectly-matched locus, and the locus was consistent with what was reported in the corresponding literature.

Integration of functional genomics data

For cell types with R-loop mapping data available, we integrated and analyzed other genomics data that might help to distinguish *bona fide* R-loops from false positives, or interpret the molecular function of R-loops and R-loop regulators (Supplementary Table S5). G4 ChIP-seq (69–71), RPA ChIP-seq (43,72), GRO-seq (13,73–76) and PRO-seq (24,77) data were searched and downloaded from GEO (78). All datasets were processed by following the original publications. ChIP-seq data for 29 different histone modifications, CLIP-seq and ChIP-seq data for validated R-loop regulators, ATAC-seq, WGBS and Repli-seq data were downloaded from ENCODE (67). R-loop-related sequence features were prepared as below. Putative G4 motifs were identified on a genome-wide scale as described previously (79). GC percent was directly downloaded from UCSC genome browser (80). AT or GC skew values were

computed for every 110 nt window with the step size of 10 nt. Predicted R-loop forming sequences were downloaded from R-loopDB (81). CrossMap was used for conversion between different genome builds whenever needed (82). Expression profiles of R-loop regulators were based on GTEx (83), TCGA (84) and CCLE (85) data.

Development of R-loopBase management system and interactive user interfaces

R-loopBase was developed using several web development technologies. Data were largely managed with MySQL. Web pages were built by HTML, CSS, JavaScript, AJAX, JSP, Java and Tomcat. The page contents were delivered by Apache. A local R-loopBase genome browser was developed on the basis of UCSC genome browser (80,86). R-loopBase is freely accessible using the URL https://rloopbase.nju.edu.cn.

RESULTS

Identification and characterization of a reference set of human R-loop zones

A major challenge in R-loop field is to precisely locate R-loops across the genome (1). Different technologies usually give rise to distinct genome-wide R-loop maps in the same cells (13,14,23–25,41,43,52,53) (Figures 1A–C). Even applying the same technology in the same cell line, different labs (42,46,51) (Figures 1A and D) or the same lab (18,21,36) (Figures 1A and E) sometimes generate R-loop maps with large differences. These broad discrepancies are largely rooted in different protocols of DNA fragmentation, R-loop enrichment and sequencing library preparation adopted by different (and even the same) technologies (Supplementary Figure S1). It is currently premature to conclude which technology or protocol is better than the others. Nevertheless, the high reproducibility between biological replicates generated in the same study indicates that each technology itself is robust if all experimental conditions are well controlled (54) (Figure 1F). With inspection, we noted that positive R-loop loci verified previously are usually well-supported by multiple independent technologies, such as, the R-loop forming region at ATRAID promoter (13) (Figure 2A). In contrast, as exemplified by SNRPN locus (87) (Figure 2A), most negative R-loop loci are not detectable or detected only by a limited number of technologies. We therefore postulate that integrative analysis of all R-loop mapping data holds the promise of distinguishing those conservative R-loop zones from technology-specific false positives or experimental variations.

As R-loop mapping technologies have been extensively applied to human cells in comparison with other species, we were therefore motivated to generate a reference set of human R-loop zones by integrative analysis. A total of 118 datasets generated in 28 human cell types under basal conditions by 11 different R-loop mapping technologies (Supplementary Figure S1) were collected and processed (Materials and Methods, Table 1 and Supplementary Table S1). Of them, 107 datasets from 26 cell types survived quality control, which took into consideration of usable reads count, reproducibility, signal enrichment and specificity and so on (Materials and Methods, Figure 2B and Supplementary Table S2). We combined all mapping data of different cell origins, aiming to characterize all possible R-loop forming regions in any human cell type. We assumed that regions detected with multiple independent technologies are more likely R-loop forming regions of high confidence by mitigating intrinsic limitations of individual technologies. Following this principle, we partitioned the R-loop regions mapped by different technologies into different confidence levels, which correspond to the minimal number of technologies they were detected by (Materials and Methods and Figure 2B). For example, the resulting R-loop zones of level 1 are those detected by $\geq$1 technology, those co-detected by $\geq$2 different technologies are classified as level 2 R-loop zones, and so on and so forth. Along with the increase of confidence levels, the number and median length of R-loop zones decreased from ~800 000 to ~200 (Figure 2C), and from ~800 nt to ~200 nt (Figure 2D), respectively. As predicted, R-loop zones of higher confidence levels were more often distributed at promoter regions (Figure 2E), and more likely associated with well-known R-loop-associated features, such as GC skew (Figure 2F), and G4 formation as determined by G4 ChIP-seq (Figure 2G), confirming that *bona fide* R-loops were increasingly enriched in R-loop zones of higher confidence levels. In support of the role of R-loops in regulating transcription termination (88), percentages of R-loops at TTS regions also gradually increased except for level 9. When compared with predicted R-loops solely based on sequence features by R-loopDB (81), a similar trend was observed (Figure 2H). Notably, a considerable fraction of high-confidence R-loops were not detected by R-loopDB (Figure 2H), suggesting that sequence feature is not the only molecular determinant for R-loop formation. In sum, we characterized a reference set of human R-loops of different confidence levels, allowing customized analysis by researchers who are interested in R-loop biology.

Compendium of human proteins regulating R-loop homeostasis

The regulatory mechanisms of R-loop dynamics have been studied and documented in scattered literature, and we were thus prompted to understand the global regulatory network by manually curating and systematically characterizing the most comprehensive list of R-loop regulatory proteins (Materials and Methods). A total of 1185 proteins in human, 24 in mouse, 63 in yeast (*Saccharomyces cerevisiae*) and 21 in *E. coli* were collected as R-loop regulators (Supplementary Table S3 and Figure 3A). We next focused on human R-loop regulators given the greater breadth of data available in human. Of 1185 human proteins, 186 (15.69%) were validated in multiple independent studies (53, 4.47%), or by multiple assays (64, 5.40%) or one assay (69, 5.82%) in only one study (Figure 3A). Among a variety of assays (Supplementary Table S3), the S9.6 antibody was most widely used (Figure 3B). However, given the specificity issue of S9.6 antibody, the sensitivity to RNase H treatment was often introduced as the gold standard (Figure 3B). It was also an important alternative to directly examine the binding, helicase or nuclease activity of a protein towards synthetic

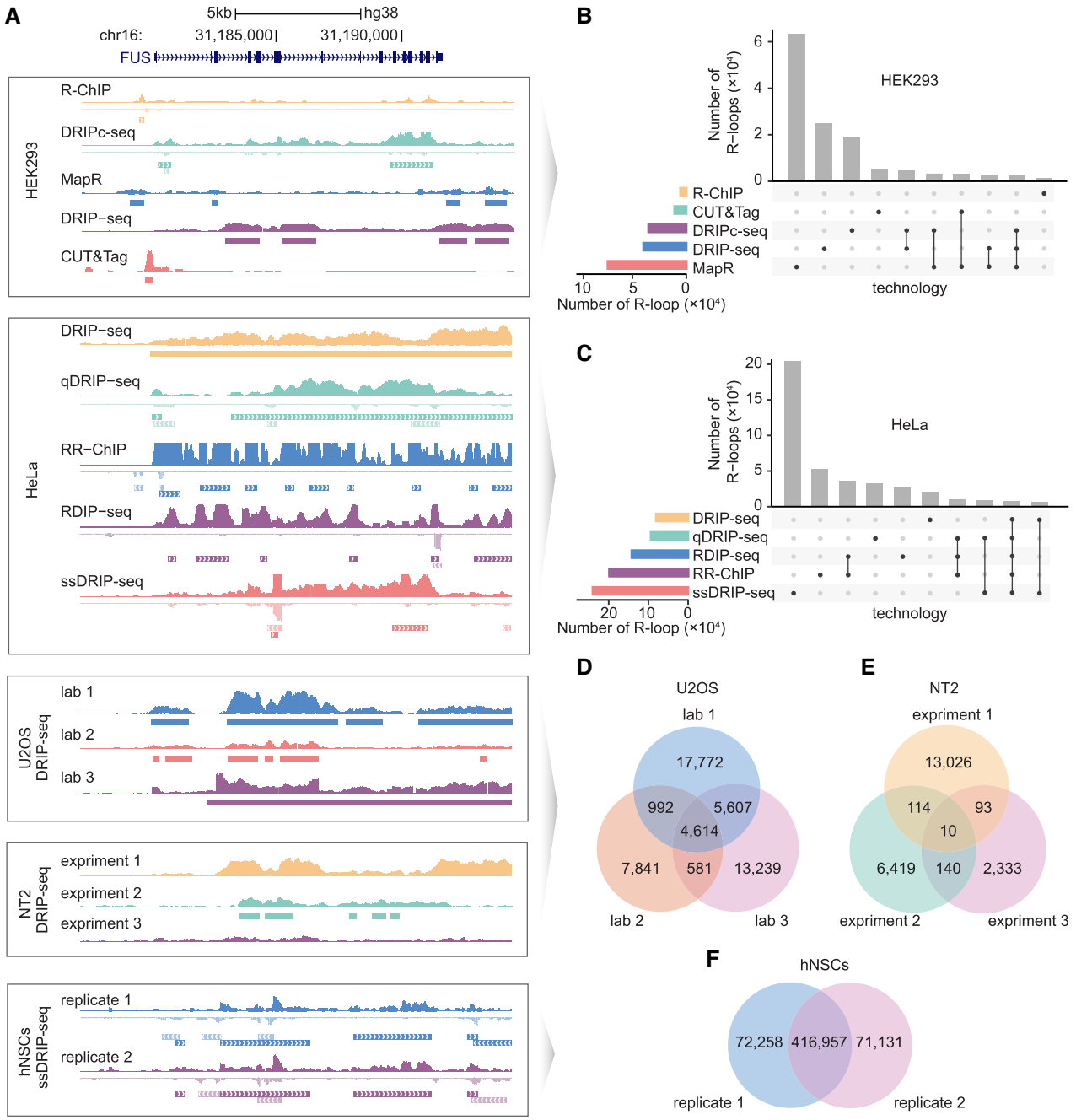

Figure 1. Broad discrepancy among mapped R-loops. (**A**) Shown are R-loop peaks and signals at *FUS* gene locus. Different technologies (the top two panels), DRIP-seq datasets generated by different labs (the third panel) and in different experiments of the same lab (the fourth panel), and two biological replicates generated by ssDRIP-seq (the bottom panel) are coded in different colors. The same color scheme is used in (B–F). (B, C) Upset plots showing the technology-specific R-loop peaks or co-detected R-loop peaks in HEK293 (**B**) or Hela (**C**) cells. (D–F) Venn diagrams showing the overlap of DRIP-seq data generated by different labs (**D**), in different experiments from the same lab (**E**), or ssDRIP-seq data generated as biological replicates in one study (**F**).

RNA:DNA hybrids *in vitro* (Figure 3B). Although not validated yet, the remaining 999 human proteins are potentially implicated in R-loop regulation as well. They were significant hits in one or two proteomics profiling studies for identification of RNA:DNA hybrid binding proteins (Figure 3A) (33,34). Similar to validated R-loop regulators, they were also enriched to be DNA, RNA or chromatin binding proteins (Supplementary Figure S3), suggesting that a considerable fraction of them may be authentic R-loop regulators. Collectively, we cataloged a comprehensive list of proteins regulating R-loops to date.

To gain an insight into how R-loop regulators are functionally related to each other, we performed PPI network analysis focusing on validated human R-loop regulators (Materials and Methods). Seven distinct clusters were readily identified, and GO enrichment analysis revealed cluster-

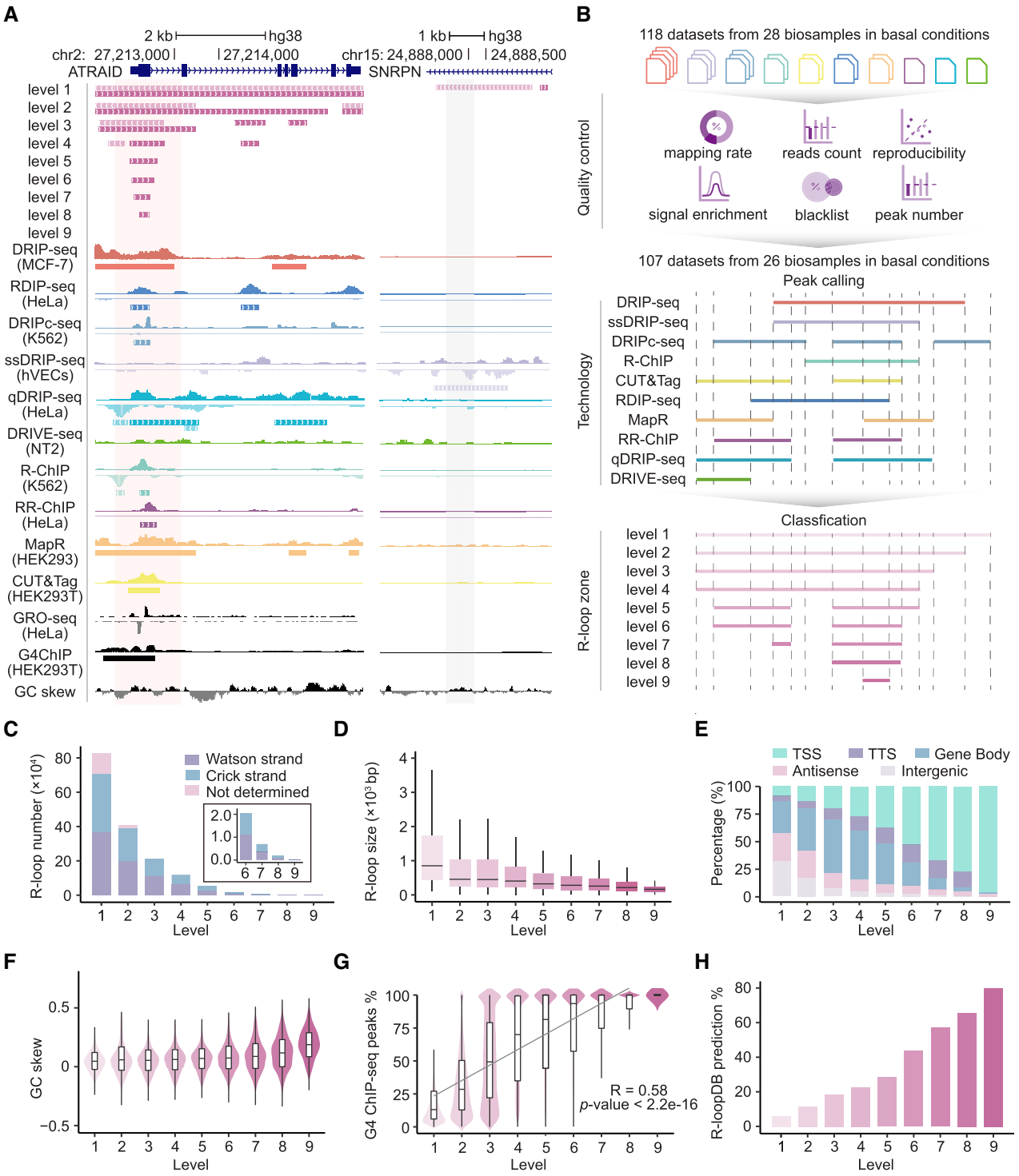

Figure 2. Identification and characterization of human R-loop zones. (**A**) Shown are two representative genomic loci covering one positive (highlighted in red rectangle) and one negative (highlighted in grey rectangle) R-loop forming regions verified previously. The R-loop signals detected by different technologies are illustrated. (**B**) Workflow for generation of human R-loop zones of different confidence levels. (**C**) Number of R-loop zones of different confidence levels. R-loop zones of levels 6–9 are further zoomed in. (D–H) Shown are R-loop size distribution (**D**), genomic distribution (**E**), GC skew (**F**), and distribution of percentages of individual R-loops overlapped with G4 ChIP-seq peaks (**G**) or predicted R-loop regions by R-loopDB (**H**) for R-loop zones of different confidence levels. TSS: transcription start sites; TTS: transcription termination sites. Linear regression is done for (**G**) and the Pearson correlation coefficient and p-value are indicated.

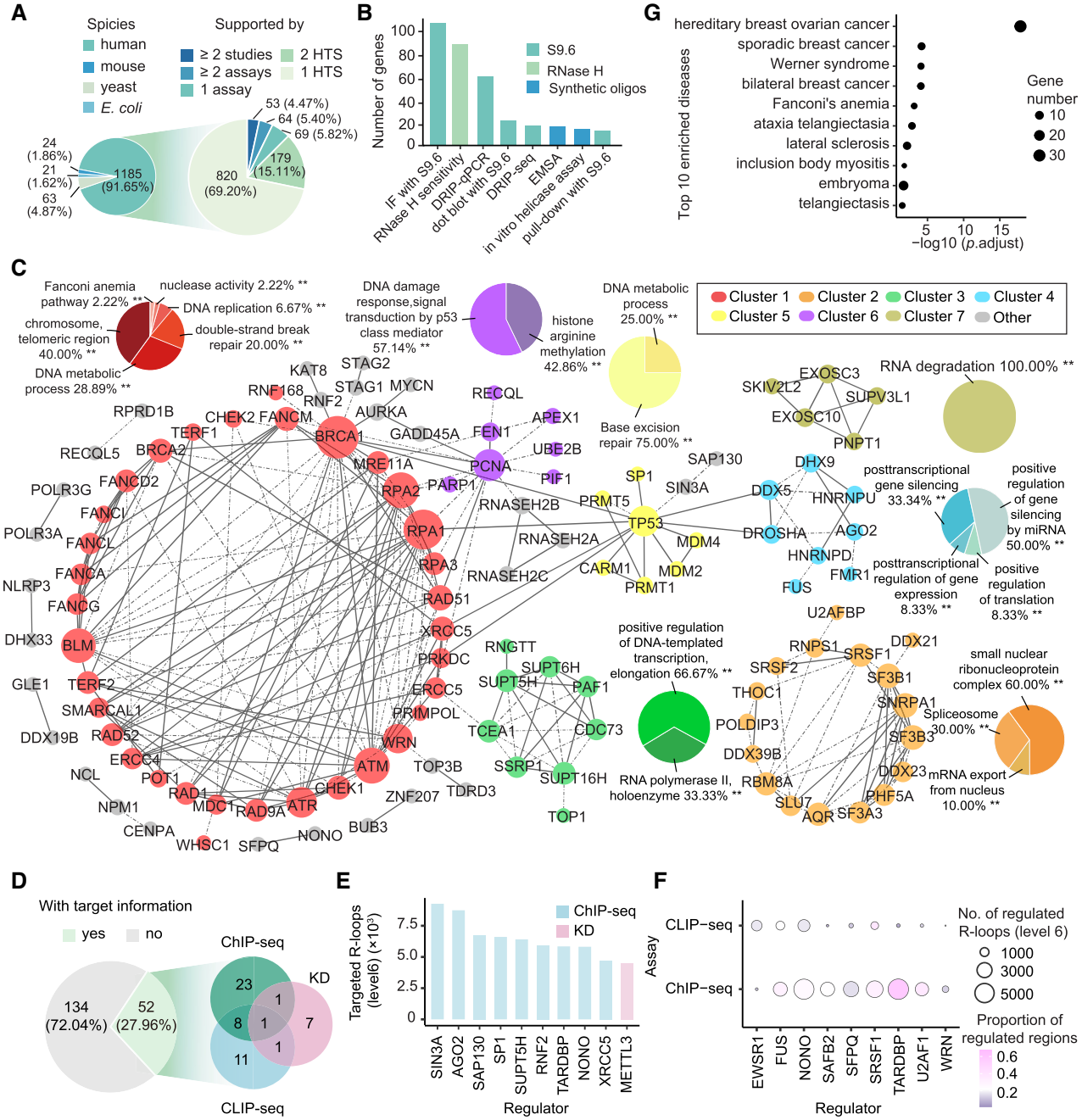

Figure 3. Characterization of R-loop regulators collected in R-loopBase. (**A**) Number of R-loop regulators in four species, and number of human R-loop regulators with different levels of supporting evidence, i.e., being validated in at least two independent studies (≥2 studies), with at least two independent assays (≥2 assays) or only one assay (1 assay), or hit in at least two high-throughput screening studies (≥2 HTS) or only one HTS study (1 HTS). (**B**) Number of top-ranked validation assays based on S9.6 monoclonal antibody, RNase H or synthetic nucleotide oligos that are indicated by different colors. (**C**) PPI network and GO enrichment analysis. PPI clusters defined by Cytoscape are labelled with different colors, and GO terms (biological process) for clusters 1–6 are shown using corresponding color schemes. Solid and dash-dot edges represent interactions of high and low confidence as defined by STRING database, respectively, and node size is proportional to interaction degree. (**D**) Left: number of regulators with or without target information. Right: number of regulators with their targeted R-loops revealed with indicated methods. (**E**) Top 10 regulators with the highest number of targeted R-loops identified by KD (pink) or ChIP-seq (blue) assay. (**F**) Shown are 9 regulators with both CLIP-seq and ChIP-seq data, with bubble size and color corresponding to number and percentage of RNA and DNA binding sites located at R-loop regions. (**G**) Bubble chart showing top 10 disease ontology (DO) terms, with the bubble size corresponding to number of genes associated with individual DO terms.

specific enrichment of biological processes (Materials and Methods and Figure 3C). Clusters 1, 5 and 6 mainly consisted of proteins involved in DNA replication or damage response processes, consistent with the deleterious role of R-loops to induce replication stress and even DNA damages as well as the regulatory role of R-loops to facilitate efficient DNA damage repair (12). Proteins in RNA metabolic processes, including RNA splicing and export factors (cluster 2) and RNA exosome complex (cluster 7), constituted the second largest group of R-loop regulators, supporting the proposed function of RNA binding proteins in counteracting R-loop formation (4). Interestingly, many RNA binding proteins (cluster 4), especially those involved in small RNA processing, are functionally connected with DNA damage repair factors, in favor of the notion that RBP re-localization upon DNA damage may coordinately regulate RNA and DNA metabolism (89). As R-loops are dynamically coupled with transcription (13), the other group of R-loop modulators mainly included transcription elongation factors (cluster 3).

With the availability of multi-omics data for human R-loop regulators, we were prompted to establish regulatory connections between individual R-loops and their regulators. The differential R-loop peaks following knockdown are likely targets of the corresponding regulator. In addition, R-loops may be directly modulated through chromatin or RNA binding activity of their regulators. Accordingly, we cataloged putative targeted R-loops for 52 validated R-loop regulators based on knockdown, ChIP-seq or CLIP-seq data (Materials and Methods and Figure 3D). Importantly, these regulatory connections were well aligned with the existing lab results. We got 175 validated R-loop forming regions targeted by one of 52 R-loop regulators (Materials and Methods and Table S4). Although the existing experimental data were mostly from cell lines different with ours, we observed good validation rates for many regulators. For example, about half of the experimentally-verified R-loop targets are well-supported for BRCA1 (69%, 9/13), SMN1 (57%, 8/14), SIN3A (50%, 4/8), U2AF1 (50%, 2/4) and etc. Therefore, the regulatory relationship we deduced here is a good starting point for mechanistic understanding of R-loop regulation. With our data, we discovered that a few proteins may be master R-loop regulators as they targeted thousands of high-confidence R-loops (Figure 3E). Nine regulators have both ChIP-seq and CLIP-seq data available, allowing us to explore their regulatory mechanism. Interestingly, although these are typical RNA binding proteins, their chromatin binding rather than RNA binding activity are in general more associated with high-confidence R-loops (Figure 3F). This finding contradicts with the view that RNA binding proteins counteract R-loop formation by binding to RNA (4), yet instead suggests RNA binding proteins may regulate gene expression and R-loop dynamics through direct association with chromatin (6,31). More follow-up mechanistic studies are thus needed to resolve the puzzle.

We further investigated the relevance of R-loop regulators to human diseases, in hope of opening new avenues for future disease mechanism research from the perspective of R-loops. Malfunction of R-loop regulators was significantly associated with human diseases. Of 186 validated human R-loop regulators, 128 are associated with human diseases, significantly higher than background (Monte Carlo simulation, P-value < 0.001). Besides cancers, neurological diseases in a broad sense, including ataxia telangiectasia and lateral sclerosis, were among the most enriched disease ontology terms (Figure 3G; Materials and Methods). It shall be interesting to interrogate why and how neural cells are specifically less tolerant of deficiency of R-loop regulators.

R-loopBase development for R-loop studies in a well-annotated genomic context

To maximize the usefulness of R-loop zones and regulators described above, we characterized them by integrating multiple categories of functional genomic annotations (Materials and Methods and Figure 4A). First, previous studies have linked R-loop formation with specific genomic features, the integration of which may help evaluate independently whether R-loop peaks detected by individual technologies or R-loop zones defined by our integrative analysis are true or not. To this end, we integrated the predicted R-loop forming sequences (81), and computed GC content, GC skew and AT skew, and predicted G4 motifs. As not all G4 motifs permit G4 formation *in vivo* and G4 structures may involve atypical motifs (90), we thus integrated G4 structures mapped with ChIP-seq technology as well. R-loop detection can rely on the identification of the displaced ssDNA. The binding profile of ssDNA-binding protein RPA was previously suggested as an alternative method to locate R-loops (4), so we also collected public RPA ChIP-seq data. R-loop formation is clearly a consequence of transcription, and histone marks can be taken as a proxy of transcription or chromatin status especially when transcription data are not available. Therefore, GRO-seq and PRO-seq data, and 29 types of histone modifications were integrated from ENCODE project (67) and elsewhere. R-loop formation is coupled with local unmethylated status of the genomic DNA (21), prompting us to further integrate whole genome bisulfide sequencing data. Hybrids between template DNA and RNA primers during DNA replication might be captured, we therefore also integrated replication timing data (91). Second, for individual R-loop regulators, we collected and organized gene annotations from HUGO, NCBI and GeneCards. We also annotated each regulator with its supporting evidences, molecular function for R-loop regulation and putative R-loop targets (Supplementary Table S3). As R-loop regulators are in general associated with human diseases (Figures 3G), to better facilitate functional study in a specific disease or cell model, we annotated their expression in normal and cancerous tissues, as well as cell lines from GTEx (83), TCGA (84) and CCLE (85) projects. Overall, about 300 datasets from ten broad categories were integrated (Figure 4A and Supplementary Table S5), generating billions of functional annotations.

To support the data management and to better serve the R-loop community, we developed R-loopBase platform with multiple user-friendly interactive interfaces (Figures 4B and 5). First, ID- and location-based query systems were developed for searching human R-loop forming regions (Figure 5A). For each query, statistics of R-loop zones

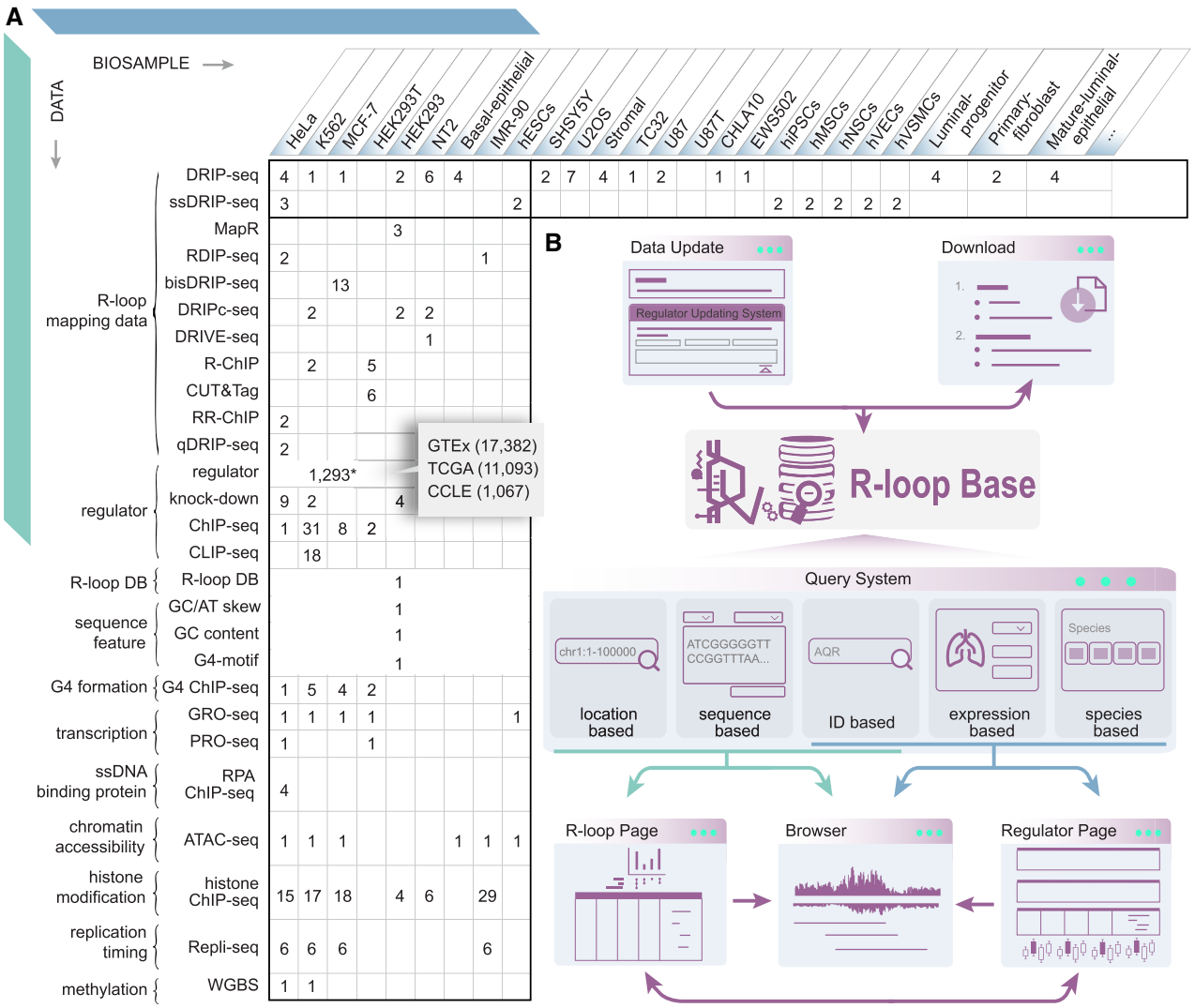

Figure 4. R-loopBase data integration and interface development. (**A**) A matrix showing the datasets integrated in R-loopBase, with columns corresponding to different cell types and rows corresponding to different types of genomics data. Data sources for expression profiles of R-loop regulators are shown in the dialog box. *Including 1,185 regulators in human, 24 in mouse, 63 in yeast and 21 in *E. coli*. (**B**) The overall framework of R-loopBase database. Datasets in (A) are systematically processed, updated and managed in R-loopBase (top left) and could be directly downloaded for genome-wide analysis (top right). In addition, through the query system (middle), all datasets could be retrieved for display on R-loop page, regulator page or genome browser (bottom).

are shown with an UpSet plot (Figure 5B). Displayed below are R-loopBase R-loop zones of different confidence levels, each of which can be accessed through the drop-down menu (Figure 5C). For each R-loop zone, the supporting technologies, cell lines in which it is detected and its known regulators are listed in detail (Figure 5C). Users will be further directed to R-loopBase genome browser via the hyperlinked genomic coordinate for visualization of R-loops in the well-annotated genomic context (Figure 5D). Alternatively, ID-, location- and sequence-based queries are supported by R-loopBase genome browser for visual inspection of R-loop zones. Clearly, along with the increase of confidence level, R-loop zones are growingly narrowed down and increasingly associated with well-known R-loop-related features, such as GC skew, G4 formation and local transcriptional activity (Figure 5D). The R-loopBase genome browser also allows direct comparison among R-loops identified by individual technologies and other genomic features.

Second, by following the hyperlink associated with a specific R-loop regulator, users can also be directed from the R-loop page to the regulator page (Figures 4B and 5C), which provides regulator-centric view of R-loopBase annotations. Besides, ID-, species- or expression-based query systems were also developed for searching R-loop regulators in specific species or with specific gene expression profile (Figure 5E). For a given R-loop regulator, four categories of gene annotations are provided (Figure 5F), i.e., basic information from public databases, supporting evidences and functions of R-loops, putative R-loop targets and gene expression profiles. We also implemented an interface for R-loopBase team members and users to independently update annotations for R-loop regulators to ensure the accu-

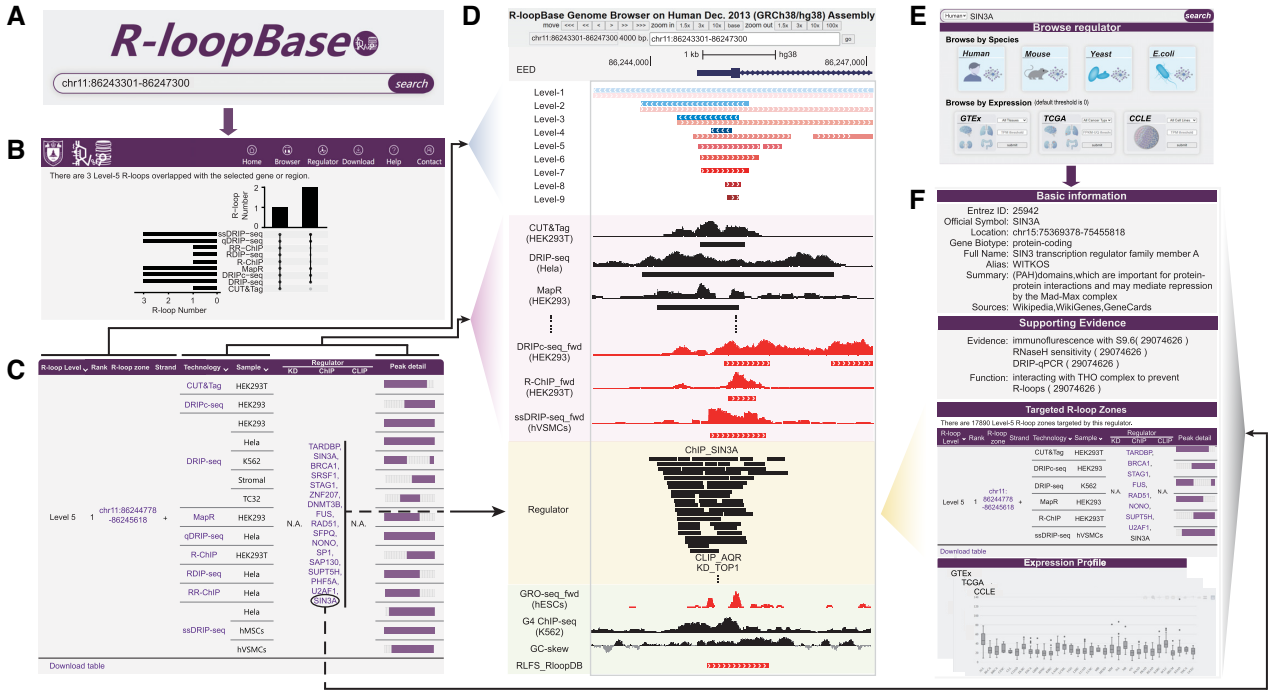

Figure 5. Screenshots of R-loopBase interfaces. (**A**) Search box for R-loops. (**B**) UpSet plot showing numbers of R-loop zones detected by indicated technologies. (**C**) Detailed information for R-loop zones. Hyperlinked texts are highlighted in purple. (**D**) A screenshot of R-loopBase genome browser showing the R-loop zone displayed in (C). Displayed from top to bottom are R-loop zones of nine levels, individual R-loop peaks mapped by different technologies, R-loop-associated features and R-loop regulators. (**E**) Query systems for R-loop regulators. (**F**) Four categories of annotation for R-loop regulator SIN3A. Cross-referenced contents are connected with arrows and shaded triangles.

racy and completeness (Figure 4B, top left). Similar to R-loop pages, regulator pages are also cross-referenced with R-loop pages and the R-loopBase genome browser. In particular, a regulator track showing DNA or RNA binding sites and regulated R-loop regions are provided on the R-loopBase genome browser. Such information may guide the future investigations of the regulatory mechanism of individual R-loop loci.

Finally, to support data analysis in batches, all searching results can be directly downloaded as tables (Figures 5C and F). Alternatively, to support genome-wide data analysis, all R-loop zones, list of regulators and targeted R-loops can be downloaded from the R-loopBase download page or genome browser.

DISCUSSION

Rapid advances in R-loop research have tremendously broadened our knowledge of R-loop regulation and function, but also led to confusions due to large discrepancies in data-driven discoveries revealed by different R-loop mapping technologies. The massive set of R-loop mapping data however has not been systematically collected and analyzed, to allow the community to easily access and compare them, thus taking full advantage of the information behind. We attempted to address this by developing R-loopBase. We integrated R-loop sequencing data in human cells generated by all 11 different technologies published to date and hundreds of R-loop-related genomics datasets. By doing so, R-loopBase functions nicely in direct comparison among different R-loop mapping technologies, and to bridge gaps between mapped R-loops at individual genomic loci and related sequence and biochemical properties. We have further endeavored to define confidence levels for mapped R-loops according to the number of technologies by which they are supported. This unique function of R-loopBase is very helpful for selecting R-loops of high confidence for functional studies. Furthermore, the regulatory mechanism for R-loop homeostasis remains elusive. By our systematic effort to collect, annotate and categorize 1293 potential R-loop regulators from literature, we provide a global view of the R-loop regulatory network for the first time. These efforts not only allow users to access regulators with various types of supporting evidence or expression profile for their own research, but also link to the regulatory mechanism and disease relevance.

To the best of our knowledge, there are two other R-loop databases, i.e., R-loopDB (81) and R-loop Atlas (http://bioinfor.kib.ac.cn/R-loopAtlas). R-loopDB is a collection of predicted R-loop forming sequences in eight species, and human and mouse R-loops mapped via DRIP-seq, DRIPc-seq or RDIP-seq. R-loop Atlas is a specialized database for R-loops in *Arabidopsis thaliana* and *Oryza sativa*, which contains R-loops mapped mainly via the ssDRIP-seq. A few key features well distinguish R-loopBase from these two databases. First, R-loopBase has the most comprehensive collection of R-loops in human. R-loopBase integrates predicted R-loops by R-loopDB, and archives human R-loop mapping datasets generated by all 11 different technologies. Second, sequence-based R-loop prediction may miss

R-loops with atypical sequence features (Figure 2H), and different R-loop mapping technologies usually generate inconsistent R-loop maps (Figure 1). It is thus not so straightforward to know the authority of a predicted or experimental R-loop from R-loopDB or R-loop Atlas. In contrast, R-loopBase presents a reference set of human R-loops and assign it with 9 different confidence levels based on integrative analysis of all R-loop mapping datasets. Moreover, R-loopBase integrates hundreds of R-loop-related genomics datasets, which collectively enable customized evaluation of the likelihood for R-loop formation at a specific genomic locus. Third, R-loop regulome is missing in either R-loopDB or R-loop Atlas. R-loopBase fills this gap by collecting and annotating a complete list of known R-loop regulatory proteins and their targeted R-loops. Users can easily get access to these resources for mechanistic and disease studies from the perspectives of R-loops.

Lastly, there is still room for further development of R-loopBase. The current release of R-loopBase is largely built on human R-loop data, mainly because only human cells have enough data for integrative analysis. However, since R-loops are conserved from bacteria to human, it would be important to include other species to study R-loops from the evolutionary point of view. While we collected R-loop regulators for three more species in addition to human, a reference set of R-loops for each species will be analyzed and presented when more genome-wide R-loop mapping data and functional genomics data are available. More importantly, as R-loops are dynamically regulated in a variety of physiological processes important for development and disease progression, our approach of defining high-confidence R-loops may miss out cell- or condition-specific R-loops under dynamic control. It is thus necessary to include R-loop dynamics data and develop novel method for precise R-loop mapping in the future. Of note, only a small fraction of R-loop regulators has been functionally validated. An even a smaller number of proteins have genomic target information available. More efforts are clearly needed to fully dissect the complex and region-specific mechanism of R-loop regulation. With continuous updates, R-loopBase will become more and more powerful as a one-stop interface to serve the community in the future.

DATA AVAILABILITY

All R-loopBase data are freely accessible using the URL: https://rloopbase.nju.edu.cn.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Chuan-Yun Li at Peking University, Lin Guo at Wuhan University and Jianhua Yang at Sun Yat-sen University for insightful suggestions for R-loopBase development and critical reading of the manuscript. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Author contributions: J.Y.C. conceived the idea. J.Y.C., L.C. and X.Z. designed the study with the help from X.D.F.. X.Z. developed the R-loopBase system and interfaces. R.L. and Y.Z analyzed the data with the help from H.G., Q.H., Z.H. and J.H.. J.Y.C. and L.C. wrote the manuscript. All authors read and approved the final manuscript.

FUNDING

J.Y.C. is supported by the National Natural Science Foundation of China [32170653]; Innovative Research Program for Overseas Returnee of Nanjing [13006002]; Start-up fund of Nanjing University [14912218]; L.C. is supported by the National Natural Science Foundation of China [32171289, 31970619]; Innovative Research Group Program of Hubei Province [2020CFA017]. Funding for open access charge: Start-up fund of Nanjing University [14912218].
Conflict of interest statement. None declared.

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.
2. 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.
3. Aguilera,A. and Garcia-Muse,T. (2012) R loops: from transcription byproducts to threats to genome stability. *Mol. Cell*, **46**, 115–124.
4. Garcia-Muse,T. and Aguilera,A. (2019) R Loops: from physiological to pathological roles. *Cell*, **179**, 604–618.
5. Richard,P. and Manley,J.L. (2017) R loops and links to human disease. *J. Mol. Biol.*, **429**, 3168–3180.
6. Chen,L., Chen,J.Y., Huang,Y.J., Gu,Y., Qiu,J., Qian,H., Shao,C., Zhang,X., Hu,J., Li,H. *et al.* (2018) The augmented R-loop is a unifying mechanism for myelodysplastic syndromes induced by high-risk splicing factor mutations. *Mol. Cell*, **69**, 412–425.
7. Tan,S.L.W., Chadha,S., Liu,Y., Gabasova,E., Perera,D., Ahmed,K., Constantinou,S., Renaudin,X., Lee,M., Aebersold,R. *et al.* (2017) A class of environmental and endogenous toxins induces BRCA2 haploinsufficiency and genome instability. *Cell*, **169**, 1105–1118.
8. Perego,M.G.L., Taiana,M., Bresolin,N., Comi,G.P. and Corti,S. (2019) R-Loops in motor neuron diseases. *Mol. Neurobiol.*, **56**, 2579–2589.
9. Lim,Y.W., Sanz,L.A., Xu,X., Hartono,S.R. and Chedin,F. (2015) Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi-Goutieres syndrome. *Elife*, **4**, e08007.
10. Posse,V., Al-Behadili,A., Uhler,J.P., Clausen,A.R., Reyes,A., Zeviani,M., Falkenberg,M. and Gustafsson,C.M. (2019) RNase H1 directs origin-specific initiation of DNA replication in human mitochondria. *PLoS Genet.*, **15**, e1007781.
11. Ouyang,J., Yadav,T., Zhang,J.M., Yang,H., Rheinbay,E., Guo,H., Haber,D.A., Lan,L. and Zou,L. (2021) RNA transcripts stimulate homologous recombination by forming DR-loops. *Nature*, **594**, 283–288.
12. Marnef,A. and Legube,G. (2021) R-loops as Janus-faced modulators of DNA repair. *Nat. Cell Biol.*, **23**, 305–313.
13. 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.
14. 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.
15. 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.
16. 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.
17. 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.

18. Sanz,L.A., Hartono,S.R., Lim,Y.W., Steyaert,S., Rajpurkar,A., Ginno,P.A., Xu,X. and Chedin,F. (2016) Prevalent, dynamic, and conserved R-loop structures associate with specific epigenomic signatures in mammals. *Mol. Cell*, **63**, 167–178.
19. Nadel,J., Athanasiadou,R., Lemetre,C., Wijetunga,N.A., P,O.B., 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.
20. Chen,P.B., Chen,H.V., Acharya,D., Rando,O.J. and Fazzio,T.G. (2015) R loops regulate promoter-proximal chromatin architecture and cellular differentiation. *Nat. Struct. Mol. Biol.*, **22**, 999–1007.
21. Ginno,P.A., Lott,P.L., Christensen,H.C., Korf,I. and Chedin,F. (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol. Cell*, **45**, 814–825.
22. 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.
23. 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.
24. 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.
25. 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.
26. Chedin,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.
27. Vanoosthuyse,V. (2018) Strengths and weaknesses of the current strategies to map and characterize R-loops. *Noncoding RNA*, **4**, 9.
28. Halasz,L., Karanyi,Z., Boros-Olah,B., Kuik-Rozsa,T., Sipos,E., Nagy,E., Mosolygo,L.A., Mazlo,A., Rajnavolgyi,E., Halmos,G. *et al.* (2017) RNA-DNA hybrid (R-loop) immunoprecipitation mapping: an analytical workflow to evaluate inherent biases. *Genome Res.*, **27**, 1063–1073.
29. 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.
30. 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.
31. Xiao,R., Chen,J.Y., Liang,Z., Luo,D., Chen,G., Lu,Z.J., Chen,Y., Zhou,B., Li,H., Du,X. *et al.* (2019) Pervasive chromatin-RNA binding protein interactions enable RNA-based regulation of transcription. *Cell*, **178**, 107–121.
32. Chakraborty,P., Huang,J.T.J. and Hiom,K. (2018) DHX9 helicase promotes R-loop formation in cells with impaired RNA splicing. *Nat. Commun.*, **9**, 4346.
33. Cristini,A., Groh,M., Kristiansen,M.S. and Gromak,N. (2018) RNA/DNA hybrid interactome identifies DXH9 as a molecular player in transcriptional termination and R-loop-associated DNA damage. *Cell Rep.*, **23**, 1891–1905.
34. Wang,I.X., Grunseich,C., Fox,J., Burdick,J., Zhu,Z., Ravazian,N., Hafner,M. and Cheung,V.G. (2018) Human proteins that interact with RNA/DNA hybrids. *Genome Res.*, **28**, 1405–1414.
35. Hartono,S.R., Malapert,A., Legros,P., Bernard,P., Chedin,F. and Vanoosthuyse,V. (2018) The aaffinity 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.
36. Ginno,P.A., Lim,Y.W., Lott,P.L., Korf,I. and Chedin,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.
37. Stork,C.T., Bocek,M., Crossley,M.P., Sollier,J., Sanz,L.A., Chedin,F., Swigut,T. and Cimprich,K.A. (2016) Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage. *Elife*, **5**, e17548.
38. Jangi,M., Fleet,C., Cullen,P., Gupta,S.V., Mekhoubad,S., Chiao,E., Allaire,N., Bennett,C.F., Rigo,F., Krainer,A.R. *et al.* (2017) SMN deficiency in severe models of spinal muscular atrophy causes widespread intron retention and DNA damage. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, E2347–E2356.
39. Zhang,X., Chiang,H.C., Wang,Y., Zhang,C., Smith,S., Zhao,X., Nair,S.J., Michalek,J., Jatoi,I., Lautner,M. *et al.* (2017) Attenuation of RNA polymerase II pausing mitigates BRCA1-associated R-loop accumulation and tumorigenesis. *Nat. Commun.*, **8**, 15908.
40. 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.
41. Manzo,S.G., Hartono,S.R., Sanz,L.A., Marinello,J., De Biasi,S., Cossarizza,A., Capranico,G. and Chedin,F. (2018) DNA topoisomerase I differentially modulates R-loops across the human genome. *Genome Biol.*, **19**, 100.
42. Abraham,K.J., Khosraviani,N., Chan,J.N.Y., Gorthi,A., Samman,A., Zhao,D.Y., Wang,M., Bokros,M., Vidya,E., Ostrowski,L.A. *et al.* (2020) Nucleolar RNA polymerase II drives ribosome biogenesis. *Nature*, **585**, 298–302.
43. Promonet,A., Padioleau,I., Liu,Y., Sanz,L., Biernacka,A., Schmitz,A.L., Skrzypczak,M., Sarrazin,A., Mettling,C., Rowicka,M. *et al.* (2020) Topoisomerase 1 prevents replication stress at R-loop-enriched transcription termination sites. *Nat. Commun.*, **11**, 3940.
44. 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.
45. Cohen,S., Puget,N., Lin,Y.L., Clouaire,T., Aguirrebengoa,M., Rocher,V., Pasero,P., Canitrot,Y. and Legube,G. (2018) Senataxin resolves RNA:DNA hybrids forming at DNA double-strand breaks to prevent translocations. *Nat. Commun.*, **9**, 533.
46. De Magis,A., Manzo,S.G., Russo,M., Marinello,J., Morigi,R., Sordet,O. and Capranico,G. (2019) DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 816–825.
47. Wu,W., Bhowmick,R., Vogel,I., Ozer,O., Ghisays,F., Thakur,R.S., Sanchez de Leon,E., Richter,P.H., Ren,L., Petrini,J.H. *et al.* (2020) RTEL1 suppresses G-quadruplex-associated R-loops at difficult-to-replicate loci in the human genome. *Nat. Struct. Mol. Biol.*, **27**, 424–437.
48. Perez-Calero,C., Bayona-Feliu,A., Xue,X., Barroso,S.I., Munoz,S., Gonzalez-Basallote,V.M., Sung,P. and Aguilera,A. (2020) UAP56/DDX39B is a major cotranscriptional RNA-DNA helicase that unwinds harmful R loops genome-wide. *Genes Dev.*, **34**, 898–912.
49. Zhang,C., Chen,L., Peng,D., Jiang,A., He,Y., Zeng,Y., Xie,C., Zhou,H., Luo,X., Liu,H. *et al.* (2020) METTL3 and N6-methyladenosine promote homologous recombination-mediated repair of DSBs by modulating DNA-RNA hybrid accumulation. *Mol. Cell*, **79**, 425–442.
50. Richard,P., Feng,S., Tsai,Y.L., Li,W., Rinchetti,P., Muhith,U., Irizarry-Cole,J., Stolz,K., Sanz,L.A., Hartono,S. *et al.* (2021) SETX (senataxin), the helicase mutated in AOA2 and ALS4, functions in autophagy regulation. *Autophagy*, **17**, 1889–1906.
51. 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.
52. Nojima,T., Tellier,M., Foxwell,J., Ribeiro de Almeida,C., Tan-Wong,S.M., Dhir,S., Dujardin,G., Dhir,A., Murphy,S. and Proudfoot,N.J. (2018) Deregulated expression of mammalian lncRNA through loss of SPT6 induces R-loop formation, replication stress, and cellular senescence. *Mol. Cell*, **72**, 970–984.
53. Yang,X., Liu,Q.L., Xu,W., Zhang,Y.C., Yang,Y., Ju,L.F., Chen,J., Chen,Y.S., Li,K., Ren,J. *et al.* (2019) m(6)A promotes R-loop formation to facilitate transcription termination. *Cell Res.*, **29**, 1035–1038.
54. Yan,P., Liu,Z., Song,M., Wu,Z., Xu,W., Li,K., Ji,Q., Wang,S., Liu,X., Yan,K. *et al.* (2020) Genome-wide R-loop landscapes during cell differentiation and reprogramming. *Cell Rep.*, **32**, 107870.
55. Edwards,D.S., Maganti,R., Tanksley,J.P., Luo,J., Park,J.J.H., Balkanska-Sinclair,E., Ling,J. and Floyd,S.R. (2020) BRD4 prevents R-loop formation and transcription-replication conflicts by ensuring efficient transcription elongation. *Cell Rep.*, **32**, 108166.

56. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
57. 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.
58. Landt,S.G., Marinov,G.K., Kundaje,A., Kheradpour,P., Pauli,F., Batzoglou,S., Bernstein,B.E., Bickel,P., Brown,J.B., Cayting,P. *et al.* (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.*, **22**, 1813–1831.
59. Gene Ontology Consortium. (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
60. Schriml,L.M., Mitraka,E., Munro,J., Tauber,B., Schor,M., Nickle,L., Felix,V., Jeng,L., Bearer,C., Lichenstein,R. *et al.* (2019) Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.*, **47**, D955–D962.
61. Wu,T., Hu,E., Xu,S., Chen,M., Guo,P., Dai,Z., Feng,T., Zhou,L., Tang,W., Zhan,L. *et al.* (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation (N Y)*, **2**, 100141.
62. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P. *et al.* (2021) The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
63. Shannon,P., Markiel,A., Ozier,O., Baliga,N.S., Wang,J.T., Ramage,D., Amin,N., Schwikowski,B. and Ideker,T. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
64. Kucera,M., Isserlin,R., Arkhangorodsky,A. and Bader,G.D. (2016) AutoAnnotate: a Cytoscape app for summarizing networks with semantic annotations. *F1000Res*, **5**, 1717.
65. Ross-Innes,C.S., Stark,R., Teschendorff,A.E., Holmes,K.A., Ali,H.R., Dunning,M.J., Brown,G.D., Gojis,O., Ellis,I.O., Green,A.R. *et al.* (2012) Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, **481**, 389–393.
66. 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.
67. Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shoresh,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A., Kaul,R. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
68. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
69. Mao,S.Q., Ghanbarian,A.T., Spiegel,J., Martinez Cuesta,S., Beraldi,D., Di Antonio,M., Marsico,G., Hansel-Hertsch,R., Tannahill,D. and Balasubramanian,S. (2018) DNA G-quadruplex structures mold the DNA methylome. *Nat. Struct. Mol. Biol.*, **25**, 951–957.
70. Zheng,K.W., Zhang,J.Y., He,Y.D., Gong,J.Y., Wen,C.J., Chen,J.N., Hao,Y.H., Zhao,Y. and Tan,Z. (2020) Detection of genomic G-quadruplexes in living cells using a small artificial protein. *Nucleic Acids Res.*, **48**, 11706–11720.
71. Lam,E.Y., Beraldi,D., Tannahill,D. and Balasubramanian,S. (2013) G-quadruplex structures are stable and detectable in human genomic DNA. *Nat. Commun.*, **4**, 1796.
72. Zhang,H., Gan,H., Wang,Z., Lee,J.H., Zhou,H., Ordog,T., Wold,M.S., Ljungman,M. and Zhang,Z. (2017) RPA Interacts with HIRA and regulates H3.3 deposition at gene regulatory elements in mammalian cells. *Mol. Cell*, **65**, 272–284.
73. Bi,M., Zhang,Z., Jiang,Y.Z., Xue,P., Wang,H., Lai,Z., Fu,X., De Angelis,C., Gong,Y., Gao,Z. *et al.* (2020) Enhancer reprogramming driven by high-order assemblies of transcription factors promotes phenotypic plasticity and breast cancer endocrine resistance. *Nat. Cell Biol.*, **22**, 701–715.
74. Core,L.J., Martins,A.L., Danko,C.G., Waters,C.T., Siepel,A. and Lis,J.T. (2014) Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nat. Genet.*, **46**, 1311–1320.
75. Estaras,C., Benner,C. and Jones,K.A. (2015) SMADs and YAP compete to control elongation of beta-catenin:LEF-1-recruited RNAPII during hESC differentiation. *Mol. Cell*, **58**, 780–793.
76. Fei,J., Ishii,H., Hoeksema,M.A., Meitinger,F., Kassavetis,G.A., Glass,C.K., Ren,B. and Kadonaga,J.T. (2018) NDF, a nucleosome-destabilizing factor that facilitates transcription through nucleosomes. *Genes Dev.*, **32**, 682–694.
77. Nilson,K.A., Lawson,C.K., Mullen,N.J., Ball,C.B., Spector,B.M., Meier,J.L. and Price,D.H. (2017) Oxidative stress rapidly stabilizes promoter-proximal paused Pol II across the human genome. *Nucleic Acids Res.*, **45**, 11088–11105.
78. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
79. Dhapola,P. and Chowdhury,S. (2016) QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res.*, **44**, W277–W283.
80. Navarro Gonzalez,J., Zweig,A.S., Speir,M.L., Schmelter,D., Rosenbloom,K.R., Raney,B.J., Powell,C.C., Nassar,L.R., Maulding,N.D., Lee,C.M. *et al.* (2021) The UCSC Genome Browser database: 2021 update. *Nucleic Acids Res.*, **49**, D1046–D1057.
81. 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.
82. Zhao,H., Sun,Z., Wang,J., Huang,H., Kocher,J.P. and Wang,L. (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, **30**, 1006–1007.
83. GTEx Consortium. (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.
84. Cancer Genome Atlas Research Network, Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R., Ozenberger,B.A., Ellrott,K., Shmulevich,I., Sander,C. and Stuart,J.M. (2013) The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.
85. Ghandi,M., Huang,F.W., Jane-Valbuena,J., Kryukov,G.V., Lo,C.C., McDonald,E.R., Barretina,J., Gelfand,E.T., Bielski,C.M., Li,H. *et al.* (2019) Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*, **569**, 503–508.
86. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
87. Bhatia,V., Barroso,S.I., Garcia-Rubio,M.L., Tumini,E., Herrera-Moyano,E. and Aguilera,A. (2014) BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature*, **511**, 362–365.
88. Skourti-Stathaki,K., Proudfoot,N.J. and Gromak,N. (2011) Human senataxin resolves RNA/DNA hybrids formed at transcriptional pause sites to promote Xrn2-dependent termination. *Mol. Cell*, **42**, 794–805.
89. Dutertre,M., Lambert,S., Carreira,A., Amor-Gueret,M. and Vagner,S. (2014) DNA damage: RNA-binding proteins protect from near and far. *Trends Biochem. Sci.*, **39**, 141–149.
90. Chambers,V.S., Marsico,G., Boutell,J.M., Di Antonio,M., Smith,G.P. and Balasubramanian,S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 877–881.
91. Pope,B.D., Ryba,T., Dileep,V., Yue,F., Wu,W., Denas,O., Vera,D.L., Wang,Y., Hansen,R.S., Canfield,T.K. *et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature*, **515**, 402–405.