Association analysis of repetitive elements and R-loop formation across species

Chao Zeng[1,*], **Masahiro Onoguchi**[1], **and Michiaki Hamada**[1,2,3,4,*]

[1]AIST-Waseda University Computational Bio Big-Data Open Innovation Laboratory (CBBD-OIL), 3-4-1, Okubo Shinjuku-ku, Tokyo 169-8555, Japan,

[2]Faculty of Science and Engineering, Waseda University, 55N-06-10, 3-4-1 Okubo Shinjuku-ku, Tokyo 169-8555, Japan

[3]Institute for Medical-oriented Structural Biology, Waseda University, 2-2, Wakamatsu-cho Shinjuku-ku, Tokyo 162-8480, Japan

[4]Graduate School of Medicine, Nippon Medical School, 1-1-5, Sendagi, Bunkyo-ku, Tokyo 113-8602, Japan

[*]Correspondence may be addressed to CZ (chao.zeng@hamadalab.com) or MH (mhamada@waseda.jp)

ABSTRACT

Genomes are known to have a large number of repetitive elements. Emerging evidence suggests that these non-coding elements may play an important regulatory role. However, few studies have investigated the effect of repetitive elements on R-loop formation. In this study, we found different repetitive elements related to R-loop formation in various species. By controlling length and genomic distributions, we observed that satellites, long interspersed nuclear elements (LINEs), and DNAs were each specifically enriched for R-loops in humans, fruit flies, and *Arabidopsis thaliana*, respectively. R-loops also tended to arise in regions of low-complexity or simple repeats across species. We also found that the repetitive elements associated with R-loop formation differ according to developmental stage. For instance, LINEs and long terminal repeats (LTRs) are more likely to contain R-loops in embryos (fruit fly) and then turn out to be low-complexity and simple repeats in post-developmental S2 cells. Our results indicate that repetitive elements may have species-specific or development-specific regulatory effects on R-loop formation. This work advances our understanding of repetitive elements and R-loop biology.

Introduction

An R-loop is a three-stranded structure composed of a DNA:RNA hybrid with a displaced single-stranded DNA (ssDNA). Although R-loops were initially considered to be rare by-products of transcription, recent reports have suggested that R-loops are widely distributed in eukaryotic genomes[1–3] and are involved in gene regulation and genome integrity[4–9]. R-loops regulate gene expression through a variety of molecular mechanisms. For instance, in promoter regions, R-loops promote or inhibit gene transcription by decreasing methylation[10] or enhancing polycomb-mediated gene silencing[11]. Surprisingly, a recent report indicated that R-loops can shield the ribosomal

gene expression by RNA polymerases II from the transcription conflicts caused by other RNA polymerases[12]. In terminator regions, R-loops can induce RNA polymerase stalling to improve the efficiency of transcription termination[13]. Simultaneously, these R-loops will promote nascent RNA cleavage and 3′ transcript degradation[14]. With respect to genome integrity, R-loops can affect genome dynamics[15] and telomere stability[16]. In centromeric regions, R-loops are reported to maintain genome stability by promoting chromatin condensation[15] or chromosome segregation[17]. In telomeric regions, telomere repeat-containing RNAs (TERRAs) preferentially accumulate in short telomeres to form telomere-repairing R-loops[16]. Previous reviews described other details of R-loops and genome integrity[4–9]. Additionally, with an increasing number of R-loop-related diseases (such as neurological diseases and cancers) being reported and studied[18–22], a deep understanding of the biology of R-loop structures is critical.

Genome-wide mapping of R-loop structures, performed with immunoprecipitation-based high-throughput sequencing[3,23], provides a global approach for quantitative analysis and systematic characterization of R-loops. This technique involves immunoprecipitating R-loop structures using a DNA-RNA hybrid antibody (S9.6) and then subjecting those R-loop-forming DNAs to direct DNA-sequencing, known as DRIP-seq (DNA–RNA immunoprecipitation followed by high-throughput DNA sequencing)[23]. Additionally, other R-loop profiling methods have been developed in the past decade by modifying antibodies and/or protocols[2,24–29].

Previous studies have uncovered several characteristics of R-loop formation. For example, within the genome, R-loops are prone to be formed in GC-skew (asymmetric strand distribution of guanines and cytosines)[23,30] or AT-skew (asymmetric strand distribution of adenines and thymines)[1,2] regions. The stability of G quadruplex (G4) structures is also related to R-loop formation[31]. Notably, multiple reports have indicated that R-loop formation is associated with short tandem repeats, especially trinucleotide repeat (TNR) expansion[32–35]. An R-loop predictive model was designed with the feature of short tandem repeats in mind[36]. Additionally, limited studies have shown that transposable elements (TEs), including TY1[37] and LINEs[38], may play a role in the formation of R-loops. Specific epigenetic signatures also connect with R-loop structures[3,38]. Although R-loops are thought to form co-transcriptionally or in cis (transcription and R-loop formation at the same locus)[6], accumulating evidence has indicated that R-loop structures can also form post-transcriptionally or in trans (RNA transcribed from a locus hybridizes to a distal locus)[39–43]. However, our understanding of R-loop formation and their sequence characteristics in the genome is still limited.

In this study, we focused on the effects of repetitive elements on R-loop formation. By comparing the frequency of repetitive elements in R-loops and the controls, we derive the repeat class or family that is associated with R-loop formation. For this purpose, we separately prepared three groups of controls, corresponding to uniform distribution in the genome, specific length and genomic distributions, and co-transcriptional formation. The three controls may not be appropriate for all the data studied; for example, we found that more than 60% of R-loops in fruit fly do not overlap with the transcribed region (implying that the third control is not preferred). We observed different

repetitive elements related to R-loop formation in various species. Based on the second control mentioned above, we discovered that satellites, LINEs, and DNAs were each enriched for R-loops in humans, fruit flies, and *Arabidopsis thaliana*, respectively. Additionally, R-loops were mainly found in regions of low-complexity or simple repeats across these three species. Interestingly, we also found that the repetitive elements associated with R-loop formation differ according to the organism's developmental stage. For fruit flies, LINEs and LTRs are more likely to contain R-loops in embryos, which changes to R-loops being more prevalent in areas of low-complexity and simple repeats in S2 cells. To our knowledge, this is the first comprehensive analysis of the association between repetitive elements and R-loop formation. This work improves our insights into the potential functions of repetitive elements and our understanding of the biological mechanisms underlying R-loop formation.

Results

Different controls provide multiple perspectives on the association between repetitive elements and R-loop formation

To investigate the association between repetitive elements and R-loop formation, we asked whether there was an overrepresented or underrepresented repetitive element in the R-loops. For this purpose, we considered three controls. First, R-loops are randomly and uniformly distributed in the genome. Second, R-loops of specific lengths are distributed in specific regions of the genome. Finally, most R-loops are co-transcriptional and are more likely to form where nascent transcripts are produced[6]. Thus, we prepared a genome control, sampling control, and GRO control to calculate the enrichment of repetitive elements in R-loops under the corresponding conditions (see "Materials and Methods" for details). We observed from the DRIP-seq data that the median R-loop lengths in animals (human and fruit fly) ranged from 414 to 618 nt, whereas those in plants (*A. thaliana*) were longer, with a median of 998 nt (Figure 1A). In addition, only 5.5%–15.8% of R-loops were detected in the intergenic region (Figure 1B), which implies that the majority of R-loops preferentially form surrounding gene regions, which is consistent with previous reports[3,24,44]. In addition, R-loops formed most frequently in promoter regions across species (Figure 1B). For *A. thaliana*, in particular, up to 60% of R-loops were distributed in promoter regions, which reduces the occurrence of R-loops in other regions of the genome, especially in introns, where only 0.2% of R-loops were formed. To this end, a sampling control was randomly generated in different datasets, simulating the distribution of R-loop lengths and genomic locations.

Considering that R-loops may form in irregular transcribed regions (e.g., enhancers, promoters) that are not detectable by ordinary RNA-seq, we used GRO-seq data that can identify nascent transcripts to define transcriptional regions (including genes) in the genome for different tissues or cell lines. We then randomly generated the GRO control from real-time transcriptional regions. Notably, according to the GRO-seq data, the proportion of nascent transcripts detected in the intergenic region varied widely among species (3.18%–10.43%, Table S1), and was highest

in humans. We speculate that the reason for this is the presence of a fraction of unannotated genes, and U2OS is a cell line that may have cancer-specific transcripts. To verify whether the majority of R-loops are co-transcriptional, we checked the percentage of overlaps between R-loops (as defined by DRIP-seq) and transcribed regions (as defined by GRO-seq). Although humans have 10.43% of transcribed regions in intergenic areas (see Table S1), the overlap between R-loops and these transcribed regions is the highest (70%) compared to the other two species (Figure 2). The lowest overlap between R-loops and transcribed regions was observed in fruit flies (embryos, 24%). It can be observed that all the transcribed regions defined in the GRO-seq data had a high match with the annotated gene regions (84.85%–91.8%, Table S1). Therefore, we speculate that the R-loops in fruit flies (embryos) may be formed by transcripts derived from distant or other chromosomal regions.

Repetitive elements are associated with R-loop formation across species and developmental stages

By first comparing the enrichment of repetitive elements in humans, fruit flies (S2 cells), and *A. thaliana*, we observed some patterns of relative consistency among species (Figure 3, middle panel). In the sampling control, we observed that low-complexity and simple repeats tended to be enriched in R-loops among species (Z-scores were greater than 10), while RC appeared less frequently in R-loops than in random cases (Z-scores were less than -1.7). Consistent with the previous report[45], DNA, LINEs, and LTRs were underrepresented in animal (human and fruit fly) R-loops (Z-scores between -2.25 and -23.28); however, they were overrepresented in plant (*A. thaliana*) R-loops (Z-scores between 0.56 and 24.9). In human and *A. thaliana* R-loops, short interspersed nuclear elements (SINEs) were underrepresented (Z-score was -18.78 and -2.73, respectively), whereas rRNA was mildly enriched (Z-score was 1.77 and 1.97, respectively). We also found some species-specific patterns. For example, satellites were significantly enriched in human R-loops (Z-score = 17.51), but significantly under-enriched in fruit fly R-loops (Z-score = -13.48). Notably, when we switched the controls to the transcriptional region (GRO control), we found that retroposons and satellites were enriched in human and *A. thaliana* R-loops, respectively, suggesting a positive correlation between the formation of cis R-loops and these two repetitive elements. Additionally, in humans and *A. thaliana*, the high degree of overlap between R-loops and transcriptional regions (Figure 2) enhances the confidence of the cis formation of R-loops in these two species.

We asked whether the association between repetitive elements and R-loop formation varies during development. For this purpose, we compared embryos and S2 cells of the fruit fly. Surprisingly, in the sampling control, we found that R-loops were more likely to form in regions of the embryo containing LINEs (Z-score = 12.22) and LTRs (Z-score = 11.86) (Figure 3, middle panel). In post-developmental S2 cells, R-loops were highly aggregated in regions with low complexity (Z-score = 11.4) or simple repeats (Z-score = 24.08). However, when we applied the GRO control (Figure 3, right panel), the above results were relatively attenuated (i.e., low-complexity and simple

repeats), as even LINEs and LTRs were not separated in the embryo (log(DRIP/GRO) = 1.06 and 1.05, respectively) and S2 cells (both log(DRIP/GRO) = 0.42) showing enrichment in the R-loops. Note that, in fruit flies, a small fraction (up to 39%) of R-loops overlapping with the transcribed regions might be a source of this inconsistency (Figure 2).

Various repeat families associated with R-loop formation

Further, we investigated the relationship between R-loop formation and the repetitive elements in each species at the repeat family level. For humans, we consistently observed an enrichment of R-loops, in the sampling and the GRO controls, containing centr (Z-score = 18.96, log(DRIP/GRO) = 1.75), telo (Z-score = 20.87, log(DRIP/GRO) = 1.53), low-complexity (Z-score = 11.34, log(DRIP/GRO) = 0.44), simple repeat (Z-score = 10.56, log(DRIP/GRO) = 0.24), snRNA (Z-score = 0.26, log(DRIP/ GRO) = 0.35) and rRNA (Z-score = 1.77, log(DRIP/GRO) = 0.47) in the genome (Figure 4). Interestingly, for the GRO control, we found that SVA repeat elements (belonging to the retroposon class, log(DRIP/GRO) = 0.54) and satellites (log(DRIP/GRO) = 0.23) preferentially occurred in the R-loops.

For fruit fly embryos, we consistently observed the enrichment of specific repeat families in R-loops in both the sampling control and the GRO control (Figure 5). For example, Gypsy (Z-score = 3.59, log(DRIP/GRO) = 0.85) and Pao (Z-score = 16.65, log(DRIP/GRO) = 1.53), which are both LTRs; Jokey (Z-score = 5.6, log(DRIP/GRO) = 1.03), R1 (Z-score = 7.28, log(DRIP/GRO) = 1.47), CR1 (Z-score = 14.35, log(DRIP/GRO) = 0.96), and LOA (Z-score = 5.97, log(DRIP/GRO) = 0.97), all belonging to the LINEs; TcMar-Tc1 (Z-score = 6.15, log(DRIP/GRO) = 0.39) and PiggyBac (Z-score = 1.95, log(DRIP/GRO) = 0.44), both of which are in the GRO control, I (log(DRIP/GRO) = 0.23) and R2 (log(DRIP/GRO) = 1.16) belonging to L1, copia (log(DRIP/GRO) = 0.92, belonging to the LTRs, and satellites (log(DRIP/GRO) = 0.53) also tended to be enriched in R-loops. For S2 cells, the number of enriched repeat families in R-loops was significantly reduced compared to that in embryos (Figure 5). In addition to the enrichment of Pao (Z-score=2.26, log(DRIP/GRO)=0.74) and CR1 (Z-score=3.45, log(DRIP/GRO)=0.32) that can still be observed in R-loops (which is consistent with the results in embryos), we also observed an increase in simple repeats (Z-score = 24.08, log(DRIP/GRO) = 0.21) and low-complexity (Z-score = 11.4, log(DRIP/GRO) = 0.15) in the R-loops of S2 cells.

For *A. thaliana*, nearly half of the repeat families were enriched in R-loops in both the sampling and GRO controls (Figure 6). These repeat families were Gypsy (Z-score = 25.7, log(DRIP/GRO) = 1.72), Copia (Z-score = 3.52, log(DRIP/GRO) = 0.99), LTR (Z-score = 8.55, log(DRIP/GRO) = 1.41), all belonging to the LTRs; MULE-MuDR (Z-score = 7.23, log(DRIP/GRO) = 1.14), CMC-EnSpm (Z-score = 7.59, log(DRIP/GRO) = 2.08), En-Spm (Z-score = 1.66, log(DRIP/GRO) = 1.52), DNA (Z-score = 0.38, log(DRIP/GRO) = 0.46), which are DNAs; L1 (Z-score = 0.56, log(DRIP/GRO) = 0.85) belonging to the LINEs; centr (Z-score = 1.34, log(DRIP/GRO) = 2.61); simple repeat (Z-score = 17.53, log(DRIP/GRO) = 0.17) and low-complexity (Z-score = 17.11, log(DRIP/GRO) =

0.22). However, PIF-Harbinger (log(DRIP/GRO) = 0.28), hAT-Ac (log(DRIP/GRO) = 0.32), all belonging to DNA, satellite (log(DRIP/GRO) = 0.58), composite (log(DRIP/GRO) = 0.53), and tRNA (log(DRIP/GRO) = 0.35), were only observed to be enriched in R-loops in the GRO control.

Discussion

R-loops and transcribed regions are not necessarily overlapping

We utilized GRO-seq data to define transcribed regions for comparison with R-loop-forming regions. This is because previous reports have shown that R-loop formation is associated with transcribed regions[3,24,44]. Unexpectedly, we found that the transcribed regions and R-loops do not inevitably overlap and differ among species and developmental stages (Figure 2). We speculate that there may be several reasons for this observation. First, different experimental conditions might result in inconsistency in the GRO-seq and DRIP-seq data, since we combined them from distinct studies. Second, as mentioned above, DRIP-seq signals that do not overlap with GRO-seq might contain transforming R-loops. Third, stable and long-lived RNAs that are not detectable with GRO-seq might form R-loops. The logic behind this may be that GRO-seq is more sensitive to nascent RNA, whereas those RNAs that have been transcribed to be stable in the R-loops are undetectable. Additionally, a large fraction of R-loops, especially in fly embryos, was derived from genic regions but did not overlap with the GRO-seq data. This might indicate that some of the genes are transcribed and stably form an R-loop in cis, suggesting that R-loop formation might not be necessarily associated with transcribed genes. These results imply that the variety of R-loop formations may be more diverse than previously expected. Further experiments with the same conditions are required to investigate these hypotheses.

Repetitive elements differently contribute to R-loop formation and function

We found that repetitive elements contribute differently to R-loop formation among the samples investigated in this study. human U2OS cells showed that 21.19% of the DRIP-seq signals contained repetitive elements, while 43.19% of the GRO-seq data contained these signatures (Figure S1A). Further analysis revealed that some repetitive sequences, especially TEs, including LINEs, SINEs, LTRs, and DNA families, did not tend to form R-loops in U2OS cells. On the other hand, low-complexity, satellite, simple repeat, retroposon, snRNA, and rRNA sequences were enriched in R-loop regions compared to non-repetitive sequences in the GRO-seq control. These results are consistent with previous reports[3,24,44,46,47]. Notably, a recent report has shown that low-complexity and simple repeat sequences are strongly associated with promoter regions[48], as are R-loop structures[3,24,44]. These results suggest that repetitive elements, such as low-complexity and simple repeats, are the key features of R-loop formation in promoter regions. Interestingly, low-complexity sequences have also been shown to be associated with Ezh2 binding, which is a component of polycomb repressive complex 2 (PRC2), and have methyltransferase activity

for histone H3 lysine 27[48]. Another report has shown that R-loop formation is required for the recruitment of PRC2 and repression of a subset of polycomb target genes[11]. These results suggest that R-loop formation involving low-complexity elements could be important for the recruitment of PRC2 and epigenetic regulation of target genes. Therefore, we hypothesize that repetitive elements in R-loop regions might contribute differently to the subsequent function of R-loop formation.

In contrast to human U2OS cells, *A. thaliana* seedlings showed that 22.25% of the DRIP-seq signals contained repetitive elements, while only 3.08% of the GRO-seq data contained these elements (Figure S1C). In addition to simple repeats, low-complexity, and satellites, which are prone to form R-loops in human U2OS cells, TEs, including LTRs, DNA transposons, and LINEs, were more preferentially enriched in R-loop regions in *A. thaliana* seedlings. These results imply that R-loop formation does not simply depend on genomic sequence features but depends highly on the species (or biological contexts). Given that R-loop formation is essential for epigenetic regulation[3,24,44], TEs that form R-loops could be critical regulatory elements for gene regulation in *A. thaliana* seedlings. Further analysis of such factors will reveal the functional significance of R-loop formation in TEs.

To investigate the contribution of repetitive elements in R-loop formation at different developmental stages, we compared the distribution of repetitive elements in R-loop regions between fly embryos and S2 cell lines. In fly embryos, 15.5% of the DRIP-seq signals contained repetitive elements, as compared to only 5.35% of the GRO-seq data (Figure S1B). In S2 cells, 9.81% of the DRIP-seq signals contained some repetitive elements, while 6.28% of the GRO-seq data contained those elements (Figure S1B). These results show that repetitive element contribution to R-loop formation is more prominent in embryos than in S2 cells, suggesting that the impact of repetitive elements on R-loop formation remarkably changes in different developmental stages or cell lineages. We also observed that LTRs, LINEs, and satellites were highly enriched in embryo R-loops and were less enriched in S2 R-loops. Conversely, simple repeats and low complexity were relatively enriched in S2 cells and less enriched in embryos. We speculate that repetitive elements could change their function through R-loop formation, along with the developmental context. For example, gypsy, which is known as one of the major insulator elements in flies[49], is more highly enriched in embryo R-loops than S2 R-loops. R-loop formation on gypsy may alter the function of the insulator or protein complex on insulator bodies, resulting in the downstream regulation of the chromatin compartment. This case is consistent with the recent observation that R-loop formation is associated with an enhancer- and insulator-like state[3]. Further investigation is required to reveal the relationship between R-loop formation and the insulator function of gypsy elements.

R-loop formation might be derived from TE regulation

Our results highlight the impact of TE elements on R-loop formation, especially at different developmental stages. This suggests that the TE sequence itself could tend to form an R-loop. Because TEs originate from exogenous

viruses, they are the target of gene silencing by multiple layers of defense mechanisms to prevent the harmful effects of TE activity. Therefore, R-loop formation involving TEs might be one such mechanism by which cells mitigate the effects of TEs. It has been shown that R-loop formation can stimulate transcription of an antisense sequence, resulting in the formation of heterochromatin[50,51]. This mechanism is suitable if R-loop formation has a role in silencing TE elements. Similarly, it is reasonable that R-loops have a role in regulatory signals of epigenetic regulations if their functional origin is derived from TE regulation. Moreover, chromatin loosening following the depletion of histone H1 induces the accumulation of R-loops in heterochromatic regions enriched with repetitive elements, including several types of TEs[52]. This result suggests that TE elements could preferentially form R-loop structures, when their silencing by heterochromatin is resolved. This is consistent with the notion that transcribing TE sequences increase the likelihood of R-loop formation. Taken together, R-loop formation might be intimately correlated with TE sequences, although further experimental studies are required to confirm this hypothesis.

Conclusion

In this study, we reanalyzed DRIP-seq data to investigate the impact of repetitive elements on R-loop formation. We found that satellites, LINEs, and DNA were enriched for R-loops in humans, fruit flies, and *A. thaliana*, respectively. Consistently, we observed that R-loops preferred to form in regions of low-complexity or simple repeats across species. Additionally, we also found that the repetitive elements associated with R-loop formation differ according to the developmental stage. LINEs and LTRs are more likely to promote R-loop formation in embryos (fruit fly). While R-loop formation changes in S2 cells to being more prevalent in low complexity and simple repeat areas of the genome. These results imply that repetitive elements may have species-specific or development-specific regulatory effects on R-loop formation. To our knowledge, this is the first study to analyze the association between repetitive elements and R-loop formation across species and developmental stages. Our results show that various repetitive elements may distinctly contribute to R-loop formation in a biological context-dependent manner. This work advances our understanding of repetitive elements and R-loop biology; future research should aim to determine the mechanism of R-loop formation on each repetitive element and its biological function.

Materials and Methods

Genomic sequences and gene annotations of human (hg38) and fruit fly (dm6) were downloaded from the UCSC genome browser[53]. The *A. thaliana* genome (TAIR10) and gene annotation were obtained from Ensembl[54]. Repetitive sequences were downloaded from the UCSC repeatmasker track[53]. We extracted genome-wide R-loop regions and transcribed regions from DRIP-seq and GRO-seq (global run-on sequencing) data, respectively, following the steps mentioned below. Bedtools (v2.25.0)[55] were used to extract (intersect sub-command) or remove (subtract sub-command) overlaps between regions. Statistics and enrichment analysis were implemented in home-made

Python scripts.

DRIP-seq analysis

Reads were aligned to the corresponding genome using BWA-MEM (0.7.17-r1188)[56] with default parameters. For paired-end reads (fruit fly and *A. thaliana*), reads aligned in a proper pair (SAMtools view -f 2)[57] were considered as mapped reads. For single-end reads (human), we extracted mapped reads (SAMtools view -F -4) for subsequent analysis. Mapped reads were sorted by SAMtools, and polymerase chain reaction (PCR) duplicates were marked with Picard MarkDuplicates (v2.18.1)[58] using default parameters. Subsequently, we discarded the read duplicates from the alignments (SAMtools view -h -F 1024).

We used DRIP-seq data from IP (immunoprecipitation), Input, and RnaseH-treated samples, separately. Using the input sample as a control, we extracted peaks in the IP and RnaseH-treated samples. MACS (v2.2.7.1) was applied to detect peaks from the alignments. In addition to the same MACS parameters (-q 0.001 –broad –broad-cutoff 0.001 –keep-dup all), we used -g hs -f BAM, -g dm -f BAMPE, and -g 1.36e8 -f BAMPE for human, fruit fly, and *A. thaliana*, respectively. We selected an output file (.broadPeak) for the final peak detection results. After removing the peaks in the IP sample that overlapped with the peaks in the RnaseH-treated sample, we obtained the final R-loop peaks.

GRO-seq analysis

For short reads with lengths of less than 100 nt (human and fruit fly), we sequentially utilized BWA-ALN and BWA-SAMSE with default parameters for mapping. Alternatively, for *A. thaliana*, we used BWA-MEM using default parameters. Homer (v4.11)[59] (findPeaks -style groseq) was applied to detect peaks (indicating transcripts). In cases where a tissue or cell line corresponded to multiple samples (e.g., biological replicates), we used BEDtools (merge sub-command) to combine peaks from all samples.

Enrichment analysis

To investigate the enrichment of repetitive elements in the R-loop-forming regions, we calculated the percentage of bases in the repetitive elements in the R-loops. For this purpose, we also prepared three control groups based on different hypotheses. First, assuming that R-loops are randomly distributed in the genome, we calculated the content of the repetitive elements in the genome as a control (referred to as “genome control”). Second, assuming that R-loops have a preferential distribution of length and genomic locations, we randomly selected 1000 groups of sequences from the genome while maintaining the same number of R-loops as well as length and genomic location (referred to as “sampling control”). Finally, assuming that R-loops are overwhelmingly formed directly where they are transcribed, we used the transcribed regions defined in the GRO-seq data as a control (referred to as “GRO control”).

For genome and GRO controls, we calculated the percentage of the bases of the repetitive elements in R-loops and control sequences x_k and y_k, respectively, and then computed the

$$enrichment_k = \log_{10}(\frac{x_k}{y_k}), \tag{1}$$

where k denotes a repetitive element (repeat class or family). For the sampling control, we computed the percentage of the bases of repetitive elements in R-loops and control groups as p_k and $\{q_{k,j}\}$ ($j \in \{1, ..., 1000\}$), respectively. Then, we calculated the Z score of p_k in $\{q_{k,j}\}$ ($j \in \{1, ..., 1000\}$) as the enrichment metric.

Availability of data and materials

DRIP-seq and GRO-seq datasets are available from public repositories. See Table S2 for the full list of the datasets.

Acknowledgements

We thank Dr. Martin Frith and Dr. Yutaka Saito for their helpful discussions. We thank Ms. Risa Maemura for her participation in the initial survey of R-loop-related studies. Computations were partially performed on the NIG supercomputer at ROIS National Institute of Genetics.

Author contributions

MH conceived and supervised this study. CZ designed and performed the experiments. CZ and MO wrote the manuscript. MH revised the manuscript critically. All authors contributed to analysis and interpretation of the data. All authors read and approved the final manuscript.

Funding

This work was supported by JSPS KAKENHI [grant numbers JP17K20032, JP16H05879, and JP20H00624 to MH; JP20K15784 to CZ].

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Wahba, L., Costantino, L., Tan, F. J., Zimmer, A. & Koshland, D. S1-drip-seq identifies high expression and polya tracts as major contributors to r-loop formation. *Genes & development* **30**, 1327–1338 (2016).
2. Xu, W. *et al.* The r-loop is a common chromatin feature of the arabidopsis genome. *Nat. plants* **3**, 704–714 (2017).

3. Sanz, L. A. *et al.* Prevalent, dynamic, and conserved r-loop structures associate with specific epigenomic signatures in mammals. *Mol. cell* **63**, 167–178 (2016).

4. Aguilera, A. & García-Muse, T. R loops: from transcription byproducts to threats to genome stability. *Mol. cell* **46**, 115–124 (2012).

5. Sollier, J. & Cimprich, K. A. Breaking bad: R-loops and genome integrity. *Trends cell biology* **25**, 514–522 (2015).

6. Chédin, F. Nascent connections: R-loops and chromatin patterning. *Trends Genet.* **32**, 828–838 (2016).

7. García-Muse, T. & Aguilera, A. R loops: from physiological to pathological roles. *Cell* **179**, 604–618 (2019).

8. Crossley, M. P., Bocek, M. & Cimprich, K. A. R-loops as cellular regulators and genomic threats. *Mol. cell* **73**, 398–411 (2019).

9. Niehrs, C. & Luke, B. Regulatory r-loops as facilitators of gene expression and genome stability. *Nat. Rev. Mol. Cell Biol.* 1–12 (2020).

10. Grunseich, C. *et al.* Senataxin mutation reveals how r-loops promote transcription by blocking dna methylation at gene promoters. *Mol. cell* **69**, 426–437 (2018).

11. Skourti-Stathaki, K. *et al.* R-loops enhance polycomb repression at a subset of developmental regulator genes. *Mol. cell* **73**, 930–945 (2019).

12. Abraham, K. J. *et al.* Nucleolar rna polymerase ii drives ribosome biogenesis. *Nature* (2020).

13. Proudfoot, N. J. Transcriptional termination in mammals: Stopping the rna polymerase ii juggernaut. *Science* **352** (2016).

14. Skourti-Stathaki, K., Proudfoot, N. J. & Gromak, N. Human senataxin resolves rna/dna hybrids formed at transcriptional pause sites to promote xrn2-dependent termination. *Mol. cell* **42**, 794–805 (2011).

15. Castellano-Pozo, M. *et al.* R loops are linked to histone h3 s10 phosphorylation and chromatin condensation. *Mol. cell* **52**, 583–590 (2013).

16. Graf, M. *et al.* Telomere length determines terra and r-loop regulation through the cell cycle. *Cell* **170**, 72–85 (2017).

17. Kabeche, L., Nguyen, H. D., Buisson, R. & Zou, L. A mitosis-specific and r loop–driven atr pathway promotes faithful chromosome segregation. *Science* **359**, 108–114 (2018).

18. Skourti-Stathaki, K. & Proudfoot, N. J. A double-edged sword: R loops as threats to genome integrity and powerful regulators of gene expression. *Genes & development* **28**, 1384–1396 (2014).

19. Groh, M. & Gromak, N. Out of balance: R-loops in human disease. *PLoS Genet.* **10**, e1004630 (2014).

20. Costantino, L. & Koshland, D. The yin and yang of r-loop biology. *Curr. opinion cell biology* **34**, 39–45 (2015).

21. Richard, P. & Manley, J. L. R loops and links to human disease. *J. molecular biology* **429**, 3168–3180 (2017).

22. Khristich, A. N. & Mirkin, S. M. On the wrong dna track: Molecular mechanisms of repeat-mediated genome instability. *J. Biol. Chem.* **295**, 4134–4170 (2020).

23. Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I. & Chédin, F. R-loop formation is a distinctive characteristic of unmethylated human cpg island promoters. *Mol. cell* **45**, 814–825 (2012).

24. Chen, L. *et al.* R-chip using inactive rnase h reveals dynamic coupling of r-loops with transcriptional pausing at gene promoters. *Mol. cell* **68**, 745–757 (2017).

25. Dumelie, J. G. & Jaffrey, S. R. Defining the location of promoter-associated r-loops at near-nucleotide resolution using bisdrip-seq. *Elife* **6**, e28306 (2017).

26. Tan-Wong, S. M., Dhir, S. & Proudfoot, N. J. R-loops promote antisense transcription across the mammalian genome. *Mol. cell* **76**, 600–616 (2019).

27. Yan, Q., Shields, E. J., Bonasio, R. & Sarma, K. Mapping native r-loops genome-wide using a targeted nuclease approach. *Cell reports* **29**, 1369–1380 (2019).

28. Malig, M., Hartono, S. R., Giafaglione, J. M., Sanz, L. A. & Chedin, F. Ultra-deep coverage single-molecule r-loop footprinting reveals principles of r-loop formation. *J. Mol. Biol.* (2020).

29. Crossley, M. P., Bocek, M. J., Hamperl, S., Swigut, T. & Cimprich, K. A. qdrip: a method to quantitatively assess rna–dna hybrid formation genome-wide. *Nucleic Acids Res.* (2020).

30. Ginno, P. A., Lim, Y. W., Lott, P. L., Korf, I. & Chédin, F. Gc skew at the $5'$ and $3'$ ends of human genes links r-loop formation to epigenetic regulation and transcription termination. *Genome research* **23**, 1590–1600 (2013).

31. De Magis, A. *et al.* Dna damage and genome instability by g-quadruplex ligands are mediated by r loops in human cancer cells. *Proc. Natl. Acad. Sci.* **116**, 816–825 (2019).

32. Lin, Y., Dent, S. Y., Wilson, J. H., Wells, R. D. & Napierala, M. R loops stimulate genetic instability of ctg· cag repeats. *Proc. Natl. Acad. Sci.* **107**, 692–697 (2010).

33. Groh, M., Lufino, M. M., Wade-Martins, R. & Gromak, N. R-loops associated with triplet repeat expansions promote gene silencing in friedreich ataxia and fragile x syndrome. *PLoS Genet.* **10**, e1004318 (2014).

34. Loomis, E. W., Sanz, L. A., Chédin, F. & Hagerman, P. J. Transcription-associated r-loop formation across the human fmr1 cgg-repeat region. *PLoS Genet.* **10**, e1004294 (2014).

35. Su, X. A. & Freudenreich, C. H. Cytosine deamination and base excision repair cause r-loop–induced cag repeat fragility and instability in saccharomyces cerevisiae. *Proc. Natl. Acad. Sci.* **114**, E8392–E8401 (2017).

36. Wongsurawat, T., Jenjaroenpun, P., Kwoh, C. K. & Kuznetsov, V. Quantitative model of r-loop forming structures reveals a novel level of rna–dna interactome complexity. *Nucleic acids research* **40**, e16–e16 (2012).

37. El Hage, A., Webb, S., Kerr, A. & Tollervey, D. Genome-wide distribution of rna-dna hybrids identifies rnase h targets in trna genes, retrotransposons and mitochondria. *PLoS Genet.* **10**, e1004716 (2014).

38. Nadel, J. *et al.* Rna: Dna hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics & chromatin* **8**, 1–19 (2015).

39. Toriumi, K., Tsukahara, T. & Hanai, R. R-loop formation in trans at an aggag repeat. *J. nucleic acids* **2013** (2013).

40. Chu, H.-P. *et al.* Terra rna antagonizes atrx and protects telomeres. *Cell* **170**, 86–101 (2017).

41. Zhou, Z., Giles, K. E. & Felsenfeld, G. Dna· rna triple helix formation can function as a cis-acting regulatory mechanism at the human β-globin locus. *Proc. Natl. Acad. Sci.* **116**, 6130–6139 (2019).

42. Ariel, F. *et al.* R-loop mediated trans action of the apolo long noncoding rna. *Mol. cell* **77**, 1055–1065 (2020).

43. Feretzaki, M. *et al.* RAD51-dependent recruitment of TERRA lncRNA to telomeres through R-loops. *Nature* (2020).

44. Yan, Q., Shields, E. J., Bonasio, R. & Sarma, K. Mapping native r-loops genome-wide using a targeted nuclease approach. *Cell reports* **29**, 1369–1380 (2019).

45. Nadel, J. *et al.* RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics & chromatin* **8**, 46 (2015).

46. Johnson, W. L. *et al.* RNA-dependent stabilization of SUV39H1 at constitutive heterochromatin. *Elife* **6** (2017).

47. Velazquez Camacho, O. *et al.* Major satellite repeat RNA stabilize heterochromatin retention of Suv39h enzymes by RNA-nucleosome association and RNA:DNA hybrid formation. *Elife* **6** (2017).

48. Lu, J. Y. *et al.* Genomic repeats categorize genes with distinct functions for orchestrated regulation. *Cell Reports* **30**, 3296 – 3311.e5 (2020).

49. Geyer, P. K. & Corces, V. G. DNA position-specific repression of transcription by a Drosophila zinc finger protein. *Genes Dev* **6**, 1865–1873 (1992).

50. Skourti-Stathaki, K., Kamieniarz-Gdula, K. & Proudfoot, N. J. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* **516**, 436–439 (2014).

51. Nakama, M., Kawakami, K., Kajitani, T., Urano, T. & Murakami, Y. DNA-RNA hybrid formation mediates RNAi-directed heterochromatin formation. *Genes Cells* **17**, 218–233 (2012).

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

53. Kent, W. J. *et al.* The human genome browser at ucsc. *Genome research* **12**, 996–1006 (2002).

54. Yates, A. D. *et al.* Ensembl 2020. *Nucleic acids research* **48**, D682–D688 (2020).

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

56. Li, H. & Durbin, R. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).

57. Li, H. *et al.* The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).

58. Picard toolkit. http://broadinstitute.github.io/picard/ (2019).

59. Wang, D. *et al.* Reprogramming transcription by distinct classes of enhancers functionally defined by erna. *Nature* **474**, 390–394 (2011).

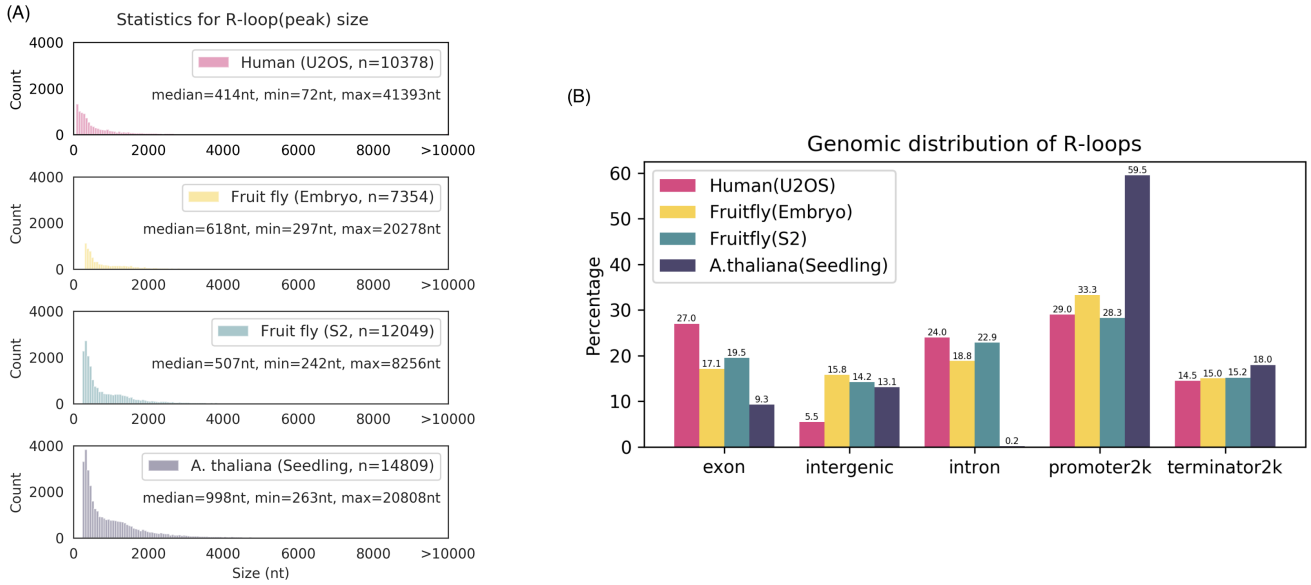

Figure 1. Distributions of R-loop characteristics in different datasets. (A) Size and (B) genomic distribution of R-loops. An R-loop is considered to be from a given genomic region when it overlaps more than one base with that region. In the case where an R-loop overlaps with more than one genomic region, we assigned it according to the priority (promoter2k > terminator2k > exon > intron > intergenic). Promoter2k: within 2000 nt upstream of a gene; terminator2k: within 2000 nt downstream of a gene; intergenic: regions excluding genes, promoter2ks, and terminator2ks.

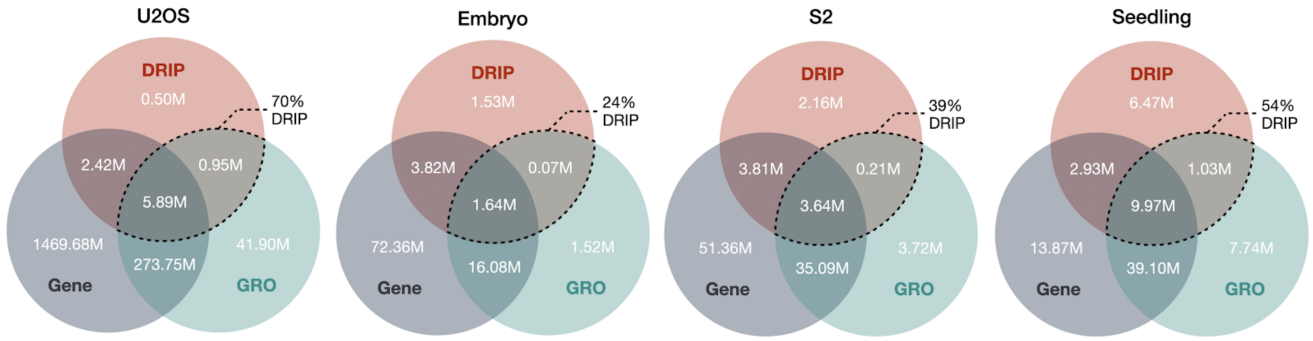

Figure 2. Venn diagram analysis for R-loops, transcribed regions, and genes at base level. R-loops and transcribed regions were defined by DRIP-seq (red) and GRO-seq (green), respectively. Gene regions (gray) were extracted based on gene annotations (see "Materials and Methods" for details). Areas marked with dashed lines indicate the overlap of R-loops and transcribed regions. The percentages represent the percentage of R-loops that are derived from the transcribed regions. M: million bases.

	log(DRIP/Genome)				Sampling(Z score)				log(DRIP/GRO)			
repClass	Human (U2OS)	Fruit fly (embryo)	Fruit fly (S2)	Arabidopsis (seedling)	Human (U2OS)	Fruit fly (embryo)	Fruit fly (S2)	Arabidopsis (seedling)	Human (U2OS)	Fruit fly (embryo)	Fruit fly (S2)	Arabidopsis (seedling)
ARTEFACT	NA	NA	NA	NA	NA	-0.29	-0.41	NA	NA	NA	NA	NA
DNA	-0.59	-0.22	-1.16	0.12	-16.68	-0.69	-9.95	8.47	-0.62	0.07	-0.82	0.98
LINE	-0.80	0.01	-0.35	0.16	-23.28	12.22	-2.25	0.56	-0.65	1.06	0.42	0.85
LTR	-0.67	-0.14	-0.60	0.25	-17.64	11.86	-8.51	24.90	-0.44	1.05	0.42	1.45
Low_complexity	0.40	-0.03	0.19	0.18	11.34	-2.00	11.40	17.11	0.44	-0.04	0.15	0.22
Other	NA	NA	NA	-0.16	NA	-2.13	-2.68	-1.62	NA	NA	NA	0.53
RC	NA	-0.71	-0.90	-0.73	-1.75	-9.07	-11.91	-15.79	NA	-0.70	-0.89	0.01
RNA	NA	-1.21	NA	NA	-2.42	-0.19	-0.48	NA	NA	-0.02	NA	NA
Retroposon	0.03	NA	NA	NA	-0.39	NA	NA	NA	0.54	NA	NA	NA
SINE	-0.08	NA	NA	-0.07	-18.78	NA	NA	-2.73	-0.23	NA	NA	0.16
Satellite	-0.14	-0.84	-1.41	-0.48	17.51	-10.10	-13.48	-1.71	1.32	0.53	-0.18	1.06
Simple_repeat	0.17	-0.12	0.24	0.13	10.56	-10.20	24.08	17.53	0.24	-0.09	0.21	0.17
Unknown	-1.14	-0.86	-1.11	NA	-3.04	-3.75	-4.96	NA	-0.90	-0.15	-0.58	NA
rRNA	0.68	-1.56	-1.82	-0.09	1.77	-2.40	-3.27	1.97	0.47	-1.45	-1.70	-0.03
scRNA	-0.15	NA	NA	NA	-2.91	NA	NA	NA	-0.45	NA	NA	NA
snRNA	0.53	NA	NA	NA	0.26	NA	NA	NA	0.35	NA	NA	NA
srpRNA	-0.01	NA	NA	NA	-2.67	NA	NA	NA	-0.20	NA	NA	NA
tRNA	0.35	NA	NA	NA	1.16	NA	NA	NA	-0.08	NA	NA	NA

Figure 3. Enrichment matrix of repeat classes across human, fruit fly and *A. thaliana*. Left, middle, and right panels correspond to genome, sampling, and GRO controls, respectively. Positive: overrepresented; negative: underrepresented; NA: not available. See Figure S1 for details.

repFamily	%Genome	%Sampling	%GRO	%DRIP	log(DRIP/Genome)	Sampling(Z score)	log(DRIP/GRO)
L1	17.06	11.78	9.89	1.39	-1.09	-21.29	-0.85
Alu	10.22	13.99	15.28	8.38	-0.09	-17.66	-0.26
ERVL-MaLR	3.62	2.67	2.11	0.64	-0.75	-14.56	-0.52
L2	3.36	3.40	4.11	1.77	-0.28	-12.76	-0.37
ERV1	2.80	2.47	1.62	0.75	-0.57	-9.30	-0.33
MIR	2.73	3.03	3.15	2.38	-0.06	-8.05	-0.12
centr	2.31	0.08	0.02	1.14	-0.31	18.96	1.75
ERVL	1.86	1.42	1.10	0.29	-0.80	-10.23	-0.57
hAT-Charlie	1.49	1.54	1.76	0.55	-0.43	-13.41	-0.51
Simple_repeat	1.23	1.29	1.05	1.83	0.17	10.56	0.24
TcMar-Tigger	1.18	1.04	1.35	0.15	-0.90	-10.20	-0.95
CR1	0.39	0.36	0.40	0.10	-0.57	-7.49	-0.59
ERVK	0.31	0.33	0.22	0.16	-0.30	-2.56	-0.15
hAT-Tip100	0.28	0.25	0.16	0.09	-0.49	-5.19	-0.24
Low_complexity	0.20	0.25	0.18	0.51	0.40	11.34	0.44
SVA	0.14	0.17	0.04	0.15	0.03	-0.39	0.54
Satellite	0.13	0.15	0.05	0.08	-0.22	-1.26	0.23
Gypsy	0.12	0.09	0.04	0.02	-0.69	-3.52	-0.28
RTE-X	0.11	0.10	0.10	0.02	-0.77	-3.92	-0.73
hAT-Blackjack	0.11	0.09	0.07	0.02	-0.71	-3.96	-0.50
TcMar-Mariner	0.09	0.08	0.07	0.01	-1.17	-3.61	-1.06
TcMar-Tc2	0.05	0.05	0.06	0.01	-0.56	-2.68	-0.59
RTE-BovB	0.04	0.04	0.04	0.01	-0.89	-3.40	-0.92
hAT	0.04	0.04	0.03	0.01	-0.52	-3.29	-0.37
Unknown	0.02	0.02	0.01	0.00	-1.14	-3.04	-0.90
MULE-MuDR	0.02	0.01	0.01	0.00	-0.68	-0.90	-0.47
tRNA-RTE	0.02	0.02	0.03	0.00	-0.75	-3.07	-0.81
LTR	0.02	0.02	0.01	0.00	NA	-2.33	NA
PiggyBac	0.02	0.02	0.02	0.01	-0.48	-1.19	-0.62
hAT-Ac	0.01	0.01	0.01	0.01	-0.16	-0.86	-0.09
tRNA	0.01	0.01	0.02	0.01	-0.01	-0.39	-0.17
Helitron	0.01	0.01	0.01	0.00	NA	-1.75	NA
snRNA	0.01	0.04	0.02	0.04	0.53	0.26	0.35
DNA	0.01	0.01	0.01	0.00	NA	-1.93	NA
telo	0.01	0.03	0.02	0.55	1.76	20.87	1.53
srpRNA	0.01	0.03	0.01	0.01	-0.01	-2.67	-0.20
5S-Deu-L2	0.01	0.01	0.01	0.00	-0.40	-1.20	-0.48
rRNA	0.01	0.02	0.01	0.04	0.68	1.77	0.47
scRNA	0.00	0.01	0.01	0.00	-0.15	-2.91	-0.45
RNA	0.00	0.01	0.00	0.00	NA	-2.42	NA
Dong-R4	0.00	0.00	0.00	0.00	NA	-0.88	NA
Penelope	0.00	0.00	0.00	0.00	-0.34	-0.75	-0.31
acro	0.00	0.00	0.00	0.00	NA	-0.26	NA
tRNA-Deu	0.00	0.00	0.00	0.00	NA	-1.35	NA
hAT-Tag1	0.00	0.00	0.00	0.00	NA	-0.64	NA
TcMar	0.00	0.00	0.00	0.00	NA	-0.63	NA
Merlin	0.00	0.00	0.00	0.00	NA	-0.34	NA
TcMar-Pogo	0.00	0.00	0.00	0.00	NA	-0.11	NA
PIF-Harbinger	0.00	0.00	0.00	0.00	NA	-0.28	NA

Figure 4. Enrichment analysis of repeat families in human U2OS cells.

		Embryo						S2					
repFamily	%Genome	%Sampling	%GRO	%DRIP	log(DRIP/Genome)	Sampling(Z score)	log(DRIP/GRO)	%Sampling	%GRO	%DRIP	log(DRIP/Genome)	Sampling(Z score)	log(DRIP/GRO)
Gypsy	7.49	3.10	0.54	3.87	-0.29	3.59	0.85	3.10	0.71	1.30	-0.76	-9.83	0.27
Simple_repeat	2.62	2.66	2.46	2.00	-0.12	-10.20	-0.09	2.72	2.75	4.51	0.24	24.08	0.21
Jockey	2.60	1.34	0.20	2.11	-0.09	5.60	1.03	1.34	0.42	1.11	-0.37	-2.17	0.42
Pao	2.16	0.99	0.10	3.34	0.19	16.65	1.53	0.99	0.22	1.21	-0.25	2.26	0.74
Satellite	1.22	1.61	0.05	0.18	-0.84	-10.10	0.53	1.60	0.07	0.05	-1.41	-13.48	-0.18
R1	0.84	0.56	0.04	1.21	0.16	7.28	1.47	0.57	0.10	0.46	-0.26	-1.62	0.66
Helitron	0.71	0.73	0.68	0.14	-0.71	-9.07	-0.70	0.77	0.69	0.09	-0.90	-11.91	-0.89
P	0.64	0.42	0.30	0.07	-0.99	-5.91	-0.66	0.43	0.24	0.05	-1.14	-7.77	-0.71
CR1	0.59	0.21	0.11	1.02	0.24	14.35	0.96	0.22	0.17	0.36	-0.21	3.45	0.32
Copia	0.52	0.25	0.03	0.22	-0.37	-0.44	0.92	0.25	0.04	0.03	-1.31	-4.95	-0.22
TcMar-Tc1	0.40	0.28	0.25	0.61	0.19	6.15	0.39	0.30	0.24	0.01	-1.54	-6.08	-1.33
Low_complexity	0.34	0.35	0.35	0.32	-0.03	-2.00	-0.04	0.36	0.37	0.52	0.19	11.40	0.15
I	0.25	0.10	0.03	0.05	-0.71	-1.18	0.23	0.10	0.02	0.06	-0.62	-1.03	0.42
LOA	0.21	0.06	0.02	0.23	0.04	5.97	0.97	0.06	0.04	0.01	-1.32	-2.12	-0.63
CMC-Transib	0.18	0.08	0.07	0.07	-0.42	-0.44	0.03	0.08	0.08	0.03	-0.80	-2.26	-0.46
Unknown	0.08	0.10	0.02	0.01	-0.86	-3.75	-0.15	0.09	0.02	0.01	-1.11	-4.96	-0.58
hAT-hobo	0.05	0.04	0.05	0.03	-0.19	-0.32	-0.14	0.04	0.02	0.00	-1.12	-2.08	-0.71
rRNA	0.05	0.05	0.04	0.00	-1.56	-2.40	-1.45	0.05	0.04	0.00	-1.82	-3.27	-1.70
Other	0.04	0.03	0.00	0.00	NA	-2.13	NA	0.03	0.00	0.00	NA	-2.68	NA
TcMar-Pogo	0.03	0.02	0.01	0.00	NA	-1.61	NA	0.02	0.01	0.00	NA	-1.94	NA
R2	0.03	0.03	0.00	0.02	-0.25	-1.08	1.16	0.03	0.00	0.00	-0.98	-2.19	0.09
RNA	0.01	0.00	0.00	0.00	-1.21	-0.19	-0.02	0.00	0.00	0.00	NA	-0.48	NA
PiggyBac	0.01	0.00	0.00	0.01	0.23	1.95	0.44	0.00	0.00	0.00	NA	-0.66	NA
MULE-NOF	0.00	0.00	0.00	0.00	NA	-0.29	NA	0.00	0.01	0.00	NA	-0.35	NA
TcMar-Mariner	0.00	0.00	0.00	0.00	NA	-0.59	NA	0.00	0.00	0.00	NA	-0.74	NA
ARTEFACT	0.00	0.00	0.00	0.00	NA	-0.29	NA	0.00	0.00	0.00	NA	-0.41	NA

Figure 5. **Enrichment analysis of repeat families in fruit fly (embryos and S2 cells).**

repFamily	%Genome	%Sampling	%GRO	%DRIP	log(DRIP/Genome)	Sampling(Z score)	log(DRIP/GRO)
Gypsy	5.21	2.98	0.18	9.44	0.26	25.70	1.72
MULE-MuDR	2.31	2.47	0.24	3.39	0.17	7.23	1.14
Helitron	1.78	1.97	0.32	0.33	-0.73	-15.79	0.01
Copia	1.43	2.05	0.25	2.43	0.23	3.52	0.99
Simple_repeat	1.18	1.23	1.08	1.60	0.13	17.53	0.17
L1	1.03	1.46	0.21	1.50	0.16	0.56	0.85
CMC-EnSpm	0.76	0.66	0.01	1.23	0.21	7.59	2.08
Satellite	0.47	0.26	0.02	0.09	-0.69	-3.28	0.58
centr	0.40	0.14	0.00	0.19	-0.32	1.34	2.61
Low_complexity	0.37	0.36	0.33	0.55	0.18	17.11	0.22
MuDR	0.27	0.29	0.03	0.52	0.28	5.24	1.29
DNA	0.27	0.21	0.07	0.22	-0.09	0.38	0.46
hAT-Ac	0.24	0.28	0.09	0.19	-0.10	-2.99	0.32
hAT	0.14	0.16	0.03	0.04	-0.57	-5.76	0.05
PIF-Harbinger	0.13	0.16	0.04	0.08	-0.20	-3.30	0.28
TcMar-Pogo	0.12	0.13	0.05	0.04	-0.48	-6.37	-0.09
En-Spm	0.09	0.09	0.00	0.13	0.16	1.66	1.52
LTR	0.06	0.05	0.00	0.13	0.30	8.55	1.41
TcMar-Stowaway	0.06	0.07	0.02	0.02	-0.45	-6.50	-0.09
SINE	0.05	0.06	0.04	0.04	-0.05	-1.63	0.06
Composite	0.05	0.05	0.01	0.04	-0.13	-1.55	0.53
tRNA	0.04	0.05	0.02	0.03	-0.08	-2.53	0.35
hAT-Tag1	0.03	0.03	0.01	0.01	-0.41	-2.25	-0.05
rRNA	0.01	0.00	0.01	0.01	-0.09	1.97	-0.03
Harbinger	0.01	0.01	0.00	0.00	-0.43	-1.08	0.02
TcMar-Mariner	0.01	0.01	0.00	0.00	-0.67	-2.27	-0.31
centromeric	0.00	0.00	0.00	0.00	NA	-0.49	NA

Figure 6. **Enrichment analysis of repeat families in *A. thaliana* seedlings.**

Dataset	#Peaks	#Bases	%Gene	%Promoter2k	%Intergenic
U2OS(DRIP)	10378	9772584	85.11	6.7	8.19
U2OS(GRO)	66659	322496237	86.71	2.86	10.43
Embryo(DRIP)	7354	7060054	77.37	5.78	16.86
Embryo(GRO)	7279	19304717	91.8	5.02	3.18
S2(DRIP)	12049	9815359	75.87	5.64	18.49
S2(GRO)	5812	42665805	90.79	4.38	4.83
Seedling(DRIP)	14809	20392020	63.26	13.57	23.17
Seedling(GRO)	12485	57831988	84.85	9.63	5.52

Table S1. Genomic distribution of DRIP- and GRO-seq peaks. Percentages of peaks overlapping with gene, promoter2k and intergenic regions shown in the last three columns. Peaks overlapping by at least 1nt with other regions are counted according to the priority (promoter2k > gene > intergenic). Promoter2k means 2000nt upstream of a gene region.

DRIP-seq			
Species	Scource	File Name	URLs
Human	U2OS	Villarreal2020_U2OS_IP_1-1	https://sra-download.ncbi.nlm.nih.gov/traces/sra64/SRR/012121/SRR12412380
		Villarreal2020_U2OS_IP_1-2	https://sra-download.ncbi.nlm.nih.gov/traces/sra51/SRR/012121/SRR12412381
		Villarreal2020_U2OS_IP_1-3	https://sra-download.ncbi.nlm.nih.gov/traces/sra58/SRR/012121/SRR12412382
		Villarreal2020_U2OS_IP_2-1	https://sra-download.ncbi.nlm.nih.gov/traces/sra77/SRR/012121/SRR12412383
		Villarreal2020_U2OS_IP_2-2	https://sra-download.ncbi.nlm.nih.gov/traces/sra57/SRR/012121/SRR12412384
		Villarreal2020_U2OS_Input-1	https://sra-download.ncbi.nlm.nih.gov/traces/sra50/SRR/012121/SRR12412385
		Villarreal2020_U2OS_Input-2	https://sra-download.ncbi.nlm.nih.gov/traces/sra57/SRR/012121/SRR12412386
		Villarreal2020_U2OS_Input-3	https://sra-download.ncbi.nlm.nih.gov/traces/sra59/SRR/012121/SRR12412387
		Villarreal2020_U2OS_RH_1-1	https://sra-download.ncbi.nlm.nih.gov/traces/sra45/SRR/012121/SRR12412388
		Villarreal2020_U2OS_RH_1-2	https://sra-download.ncbi.nlm.nih.gov/traces/sra63/SRR/012121/SRR12412389
		Villarreal2020_U2OS_RH_2-1	https://sra-download.ncbi.nlm.nih.gov/traces/sra71/SRR/012121/SRR12412390
		Villarreal2020_U2OS_RH_2-2	https://sra-download.ncbi.nlm.nih.gov/traces/sra55/SRR/012121/SRR12412391
		Villarreal2020_U2OS_RH_2-3	https://sra-download.ncbi.nlm.nih.gov/traces/sra38/SRR/012121/SRR12412392
Fruit fly	Embryo	Alecki2020_Embryo_IP_1	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443751/SRR8645632/SRR8645632.sra
		Alecki2020_Embryo_IP_2	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443752/SRR8645633/SRR8645633.sra
		Alecki2020_Embryo_RH_1	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443753/SRR8645634/SRR8645634.sra
		Alecki2020_Embryo_RH_2	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443754/SRR8645635/SRR8645635.sra
		Alecki2020_Embryo_Input_1	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443755/SRR8645636/SRR8645636.sra
		Alecki2020_Embryo_Input_2	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443756/SRR8645637/SRR8645637.sra
	S2	Alecki2020_S2_IP_1	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443763/SRR8645644/SRR8645644.sra
		Alecki2020_S2_IP_2	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443764/SRR8645645/SRR8645645.sra
		Alecki2020_S2_RH_1	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443765/SRR8645646/SRR8645646.sra
		Alecki2020_S2_RH_2	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443766/SRR8645647/SRR8645647.sra
		Alecki2020_S2_Input_1	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443767/SRR8645648/SRR8645648.sra
		Alecki2020_S2_Input_2	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX544/SRX5443768/SRR8645649/SRR8645649.sra
A.thaliana	seedlings	Xu2017_Seedling_IP_1	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX261/SRX2617507/SRR5318040/SRR5318040.sra
		Xu2017_Seedling_IP_2	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX261/SRX2617508/SRR5318041/SRR5318041.sra
		Xu2017_Seedling_RH_1	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX261/SRX2617509/SRR5318042/SRR5318042.sra
		Xu2017_Seedling_RH_2	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX261/SRX2617510/SRR5318043/SRR5318043.sra
		Xu2017_Seedling_Input	ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dra/sralite/ByExp/litesra/SRX/SRX286/SRX2867541/SRR5626994/SRR5626994.sra

GRO-seq			
Human	U2OS	GRO_U2OS	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66928
Fruit fly	Embryo	GRO_Embryo	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE41611
	S2	GRO_S2_A	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE68677
		GRO_S2_B	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE23543
A.thaliana	seedlings	GRO_seedlings	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE117014

Table S2. DRIP-seq and GRO-seq datasets used in this study.

(A)

repClass	%Genome	%Sampling	%GRO	%DRIP	log(DRIP/Genome)	Sampling(Z score)	log(DRIP/GRO)
LINE	20.96	15.67	14.56	3.30	-0.80	-23.28	-0.65
SINE	12.99	17.06	18.48	10.77	-0.08	-18.78	-0.23
LTR	8.80	7.05	5.15	1.89	-0.67	-17.64	-0.44
DNA	3.33	3.16	3.56	0.86	-0.59	-16.68	-0.62
Satellite	2.46	0.26	0.09	1.78	-0.14	17.51	1.32
Simple_repeat	1.23	1.29	1.05	1.83	0.17	10.56	0.24
Low_complexity	0.20	0.25	0.18	0.51	0.40	11.34	0.44
Retroposon	0.14	0.17	0.04	0.15	0.03	-0.39	0.54
Unknown	0.02	0.02	0.01	0.00	-1.14	-3.04	-0.90
RC	0.01	0.01	0.01	0.00	NA	-1.75	NA
snRNA	0.01	0.04	0.02	0.04	0.53	0.26	0.35
srpRNA	0.01	0.03	0.01	0.01	-0.01	-2.67	-0.20
rRNA	0.01	0.02	0.01	0.04	0.68	1.77	0.47
scRNA	0.00	0.01	0.01	0.00	-0.15	-2.91	-0.45
tRNA	0.00	0.01	0.01	0.01	0.35	1.16	-0.08
RNA	0.00	0.01	0.00	0.00	NA	-2.42	NA
Total	50.19	45.06	43.19	21.19			

(B)

		Embryo						S2					
repClass	%Genome	%Sampling	%GRO	%DRIP	log(DRIP/Genome)	Sampling(Z score)	log(DRIP/GRO)	%Sampling	%GRO	%DRIP	log(DRIP/Genome)	Sampling(Z score)	log(DRIP/GRO)
LTR	10.18	4.33	0.67	7.43	-0.14	11.86	1.05	4.35	0.97	2.54	-0.60	-8.51	0.42
LINE	4.52	2.30	0.40	4.63	0.01	12.22	1.06	2.31	0.76	2.00	-0.35	-2.25	0.42
Simple_repeat	2.62	2.66	2.46	2.00	-0.12	-10.20	-0.09	2.72	2.75	4.51	0.24	24.08	0.21
DNA	1.31	0.85	0.67	0.79	-0.22	-0.69	0.07	0.87	0.60	0.09	-1.16	-9.95	-0.82
Satellite	1.22	1.61	0.05	0.18	-0.84	-10.10	0.53	1.60	0.07	0.05	-1.41	-13.48	-0.18
RC	0.71	0.73	0.68	0.14	-0.71	-9.07	-0.70	0.77	0.69	0.09	-0.90	-11.91	-0.89
Low_complexity	0.34	0.35	0.35	0.32	-0.03	-2.00	-0.04	0.36	0.37	0.52	0.19	11.40	0.15
Unknown	0.08	0.10	0.02	0.01	-0.86	-3.75	-0.15	0.09	0.02	0.01	-1.11	-4.96	-0.58
rRNA	0.05	0.05	0.04	0.00	-1.56	-2.40	-1.45	0.05	0.04	0.00	-1.82	-3.27	-1.70
Other	0.04	0.03	0.00	0.00	NA	-2.13	NA	0.03	0.00	0.00	NA	-2.68	NA
RNA	0.01	0.00	0.00	0.00	-1.21	-0.19	-0.02	0.00	0.00	0.00	NA	-0.48	NA
ARTEFACT	0.00	0.00	0.00	0.00	NA	-0.29	NA	0.00	0.00	0.00	NA	-0.41	NA
Total	21.06	13.01	5.35	15.50				13.16	6.28	9.81			

(C)

repClass	%Genome	%Sampling	%GRO	%DRIP	log(DRIP/Genome)	Sampling(Z score)	log(DRIP/GRO)
LTR	6.70	5.08	0.43	11.99	0.25	24.90	1.45
DNA	4.42	4.56	0.62	5.86	0.12	8.47	0.98
RC	1.78	1.97	0.32	0.33	-0.73	-15.79	0.01
Simple_repeat	1.18	1.23	1.08	1.60	0.13	17.53	0.17
LINE	1.03	1.46	0.21	1.50	0.16	0.56	0.85
Satellite	0.87	0.40	0.03	0.29	-0.48	-1.71	1.06
Low_complexity	0.37	0.36	0.33	0.55	0.18	17.11	0.22
SINE	0.09	0.11	0.05	0.08	-0.07	-2.73	0.16
Other	0.05	0.05	0.01	0.04	-0.16	-1.62	0.53
rRNA	0.01	0.00	0.01	0.01	-0.09	1.97	-0.03
Total	16.51	15.22	3.08	22.25			

Figure S1. Enrichment analyses of repeat classes across (A)human, (B)fruit fly, and (C)*A. thaliana*.