PNAS RESEARCH ARTICLE | GENETICS OPEN ACCESS

Triple-helix potential of the mouse genome

Kaku Maekawa[a,b], Shintaro Yamada[a,b], Rahul Sharma[c], Jayanta Chaudhuri[c], and Scott Keeney[a,d,1]

Contributed by Scott Keeney; received March 7, 2022; accepted March 30, 2022; reviewed by Alain Nicolas and Lorraine Symington

Certain DNA sequences, including mirror-symmetric polypyrimidine•polypurine runs, are capable of folding into a triple-helix–containing non–B-form DNA structure called H-DNA. Such H-DNA–forming sequences occur frequently in many eukaryotic genomes, including in mammals, and multiple lines of evidence indicate that these motifs are mutagenic and can impinge on DNA replication, transcription, and other aspects of genome function. In this study, we show that the triplex-forming potential of H-DNA motifs in the mouse genome can be evaluated using S1-sequencing (S1-seq), which uses the single-stranded DNA (ssDNA)–specific nuclease S1 to generate deep-sequencing libraries that report on the position of ssDNA throughout the genome. When S1-seq was applied to genomic DNA isolated from mouse testis cells and splenic B cells, we observed prominent clusters of S1-seq reads that appeared to be independent of endogenous double-strand breaks, that coincided with H-DNA motifs, and that correlated strongly with the triplex-forming potential of the motifs. Fine-scale patterns of S1-seq reads, including a pronounced strand asymmetry in favor of centrally positioned reads on the pyrimidine-containing strand, suggested that this S1-seq signal is specific for one of the four possible isomers of H-DNA (H-y5). By leveraging the abundance and complexity of naturally occurring H-DNA motifs across the mouse genome, we further defined how polypyrimidine repeat length and the presence of repeat-interrupting substitutions modify the structure of H-DNA. This study provides an approach for studying DNA secondary structure genome-wide at high spatial resolution.

non-B DNA | H-DNA | DNA topology | triplex DNA | genomics

DNA sequences that can form non–B-form structures occur frequently in noncoding regions of eukaryotic genomes (1, 2). Several lines of evidence show that these motifs have biological impacts (3–6). However, their non-B DNA-forming potentials are not fully understood.

One such structure, H-DNA, consists of an intramolecular triplex plus single-stranded DNA (ssDNA) (Fig. 1*A*) (9). The duplex from one half of the H-DNA motif forms Hoogsteen triplets with a "third strand" contributed by melting the duplex in the other half of the motif. Polypyrimidine•polypurine mirror repeats are prominent examples of H-DNA motifs. H-DNA contains three ssDNA regions that we will refer to as the central loop, orphan strand, and junction (Fig. 1*A*). The central loop is an ssDNA hairpin between the triplex-forming segments on the same strand that contributes the third strand to the triplex. The junction is a short ssDNA region between the triplex and the flanking duplex. The orphan strand is complementary to the central loop, the third strand, and the junction.

H-DNA motifs are enriched in genomes of yeast, human, and other species (10, 11). H-DNA motifs are overrepresented at introns and promoters (12, 13), coding regions of several disease-involved genes (14), transposable elements (15), and fragile sites and oncogenic translocation breakpoints (16, 17). TC repeats, one type of H-DNA motif, are common in mammals (18) and in 5′ noncoding regions of plant genomes (19).

Intramolecular triplexes impinge on transcription, replication, and recombination (9, 20). For example, $(GAA)_n$ repeats stall replication in vivo (21) and block transcription (22, 23). So-called "suicidal" mirror repeats can arrest DNA polymerase in vitro (24), and H-DNA–forming sequences at the *c-MYC* locus interfere with transcription (25, 26).

H-DNA motifs are mutagenic (27). Triplex formation promotes genetic instability, mutation, and recombination leading to repeat expansion or genomic rearrangement (3). Transgenic mouse models indicate that H-DNA can induce genetic instability (28), while H-DNA motifs in the human genome are correlated with increased frequencies of somatic mutations, including recurrent mutations (17).

H-DNA in eukaryotic genomes has been analyzed by various methods. Recognition of nuclei by triplex-specific monoclonal antibodies (29, 30) could be competed by exogenous triplex DNA (31). Staining of interphase human cell nuclei with triplex-specific antibodies overlapped with sites of hybridization with probes for the displaced

Significance

H-DNA is a non–B-form DNA structure containing intramolecular triplex and single-stranded DNA (ssDNA) regions. H-DNA–forming motifs include polypyrimidine•polypurine mirror repeat sequences, which occur frequently in eukaryotic genomes. These motifs have biological impacts on genome stability and processes such as replication and transcription, but their non-B DNA-forming potentials are not fully understood. Here, we show that the triplex-forming potential of H-DNA motifs in the mouse genome can be evaluated by a deep-sequencing technology that uses the ssDNA-specific nuclease S1 to detect ssDNA. It is currently unclear whether the H-DNA detected formed in vitro or was already present in vivo. Nevertheless, this study provides an approach to unveiling structural features of intramolecular triplexes genome-wide at high spatial resolution.

Author affiliations: [a]Molecular Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065; [b]Department of Radiation Genetics, Graduate School of Medicine, Kyoto University, Kyoto 606-8501, Japan; [c]Immunology Program, Memorial Sloan Kettering Cancer Center, New York, NY 10065; and [d]HHMI, Memorial Sloan Kettering Cancer Center, New York, NY 10065

Author contributions: K.M., S.Y., and S.K. designed research; K.M., S.Y., and R.S. performed research; R.S. and J.C. contributed new reagents/analytic tools; K.M., S.Y., and S.K. analyzed data; and K.M., S.Y., and S.K. wrote the paper.

Reviewers: A.N., Institut Curie; and L.S., Columbia University Irving Medical Center.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: s-keeney@ski.mskcc.org.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2203967119/-/DCSupplemental.

Published May 3, 2022.

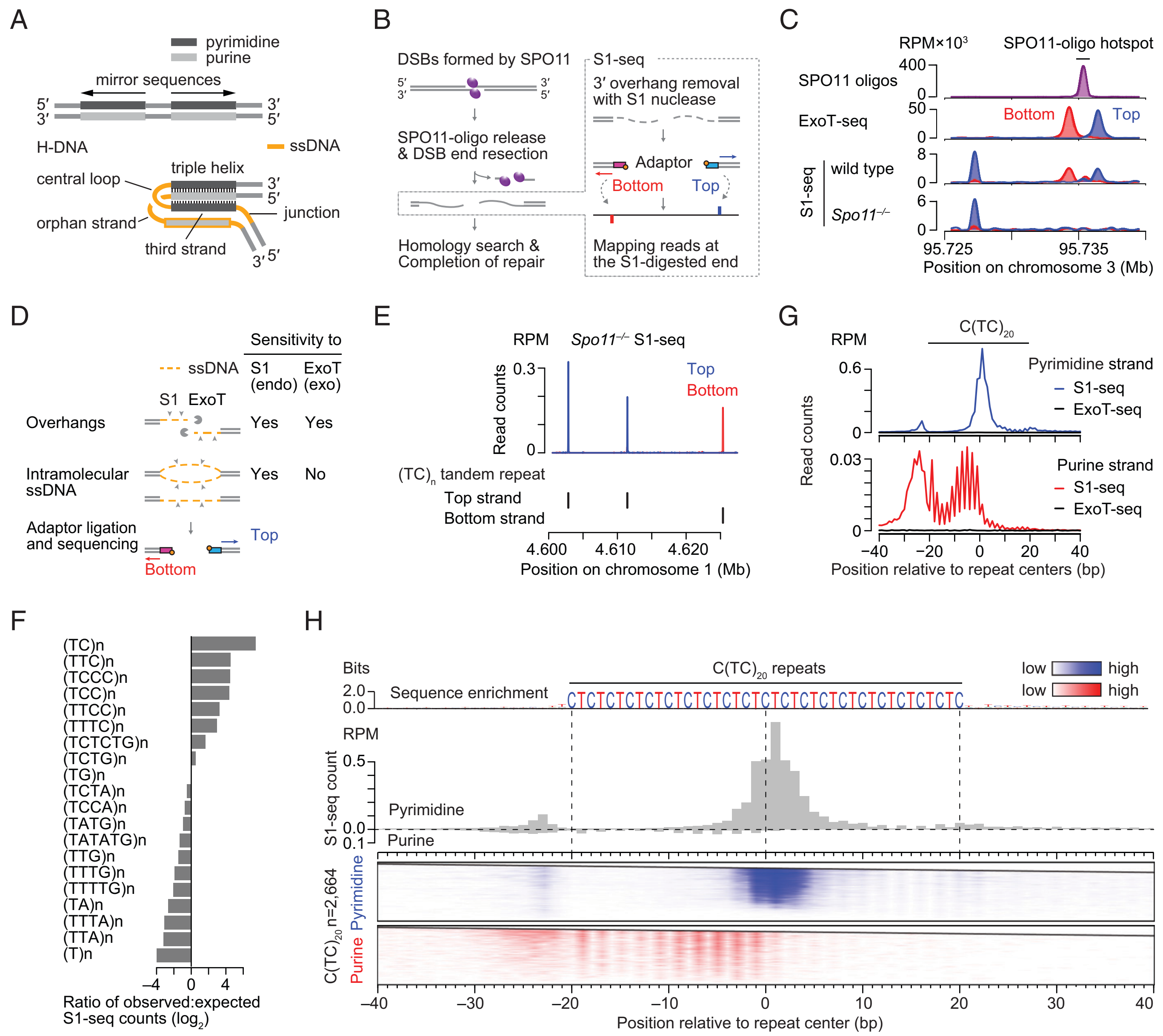


Fig. 1. Spo11-independent S1-seq clusters at H-DNA motifs. (*A*) Schematic of H-DNA, consisting of a triplex and ssDNA, formed on a polypyrimidine•polypurine mirror sequence. (*B*) Early steps in meiotic recombination and overview of the S1-seq method. SPO11 (magenta ellipses) cuts DNA via a covalent protein–DNA intermediate. In S1-seq, sequencing adaptors are linked to duplex ends generated by removal of ssDNA tails using nuclease S1. Throughout, top strand refers to the strand that runs 5′ to 3′ in the genome assembly. The two adaptors are identical in sequence but are color coded blue or red to indicate whether the read maps to the top or bottom strand, respectively. (*C*) Strand-specific S1-seq (reads per million mapped [RPM]) at a representative DSB hot spot (right side, coincident with a peak in the SPO11-oligo sequencing) with a SPO11-independent read cluster nearby (left side). S1-seq and ExoT sequencing are from ref. 7. SPO11-oligo sequencing data are from ref. 8. (*D*) Substrate specificities for nuclease S1 and ExoT. (*E*) Examples of pyrimidine-strand SPO11-independent S1-seq clusters at TC repeats. Black ticks below the plot show annotated TC repeats (RepeatMasker) on the top or bottom strand. (*F*) Preferential enrichment of S1-seq reads from *Spo11*$^{-/-}$ mice at a subset of pyrimidine repeats (RepeatMasker annotations). (*G*) Averaged S1-seq and ExoT-seq signal around C(TC)$_{20}$ sequences (*n* = 2,664). Note the different *y* axis scales for pyrimidine- vs. purine-strand reads. (*H*) Stereotyped S1-seq read distributions at C(TC)$_{20}$ sequences (*n* = 2,664). The sequence logo indicates the base frequency. The histogram (gray) shows the absolute average read count, illustrating the strong strand asymmetry. The heat maps below show the S1-seq reads separated by strand (pyrimidine strand in blue; purine strand in red). Each line is a single C(TC)$_{20}$ sequence, ranked from highest total S1-seq read count at the top. The color gradients are calibrated separately for each strand to facilitate displaying the spatial patterning for the weaker S1-seq signal on the purine strand.

ssDNA (32). The dye Thiazole Orange, which binds to triplex DNA in vitro, also binds to dipteran chromosomal regions that can be labeled by anti-triplex antibodies (33).

Unanswered questions include which motifs form H-DNA, and what are the specific triplex structures formed? A significant obstacle is the lack of methods to visualize H-DNA genome-wide at high resolution. Using potassium permanganate as a chemical probe for ssDNA demonstrated that there is a propensity toward single-strandedness near H-DNA motifs in vivo (34). However, it is unclear whether the ssDNA detected was indeed H-DNA and, if so, what the structure of the H-DNA was, especially at individual loci.

We examined the triplex-forming potential of H-DNA motifs in mouse genomic DNA using S1-sequencing (S1-seq). In S1-seq, high–molecular weight genomic DNA is embedded in agarose and digested with the ssDNA-specific nuclease S1, then adaptors are ligated to the blunted DSB ends for deep sequencing (7, 35, 36) (Fig. 1*B*). We show that S1-seq applied to DNA from mouse testis detects a prominent signal at H-DNA motifs, and we use this signal to study structural features of intramolecular triplexes.

Results

SPO11-Independent S1-seq Signal. During meiosis, DNA double-strand breaks (DSBs) formed by SPO11 are exonucleolytically processed to generate ssDNA tails that engage in homologous recombination (37–39). We previously developed S1-seq to study this DSB resection (7, 35, 36). When applied to mouse testis samples, S1-seq reads were enriched near meiotic DSB hot spots in wild-type C57BL/6J (B6) mice relative to *Spo11*$^{-/-}$ mice (Fig. 1*C*) (7). Because DNA ends on both sides of a DSB are resected, S1-seq signal mapping to the top strand spreads to the right from DSB hot spots, while bottom strand reads spread to the left.

However, we also observed S1-seq reads that were reproducibly enriched at sites distinct from meiotic DSB hot spots, that were similarly enriched in mice lacking SPO11, and that showed a strong strand bias (Fig. 1*C* and *SI Appendix*, Fig. S1). Little or no signal was seen at these sites if genomic DNA was digested with the 3′ → 5′ exonuclease ExoT (ExoT-seq) instead of the endonuclease S1, unlike around meiotic DSB hot spots (Fig. 1*C* and *SI Appendix*, Fig. S1). ExoT-seq is specific for DSBs (7, 40), but the endonuclease S1 can generate ligatable junctions at other ssDNA-containing structures in addition to DSBs, such as bubbles, nicks, or gaps (Fig. 1*D*). Therefore, we inferred from the lack of signal from ExoT-seq that SPO11-independent S1-seq signal arises from a non-DSB structure(s) rather than from SPO11-independent DSBs such as those associated with replication or transposon activity.

In searching for features shared among DSB-independent S1-seq clusters (DISCs), we noticed that many were located at polypyrimidine repeats with the most prominent enrichment at TC repeats and with a strong strand asymmetry in which most reads mapped to the pyrimidine strand (Fig. 1*E*). When S1-seq read density at short tandem repeats in *Spo11*$^{-/-}$ mice was calculated to evaluate this correlation, enrichment was observed for a subset of polypyrimidine repeats, including TC repeats (Fig. 1*F*).

S1-seq showed a stereotyped signal distribution around TC repeats. To illustrate this, Fig. 1 *G* and *H* show the signal around all C(TC)$_{20}$ repeats. [We chose C(TC)$_{20}$ for this example because there are many copies in the mouse genome and because it is mirror-symmetric.] On the pyrimidine strand, S1-seq showed a major cluster of reads at the repeat center, a minor cluster immediately to the left of the repeat, and a weak striated signal in the right half of the repeat. In contrast, the purine strand showed a weak, broad, striated enrichment within the left half of the repeat and a weak, diffuse signal in the sequence flanking the repeat on the left. We concluded that S1-seq likely detects intramolecular ssDNA preferentially formed in a sequence-dependent manner at pyrimidine-rich repeats. In principle, the ssDNA could have been present in vivo, but it is also possible that it mostly was formed ex vivo, when the genomic DNA was in agarose plugs (discussed further in *S1-seq Patterns at H-DNA Motifs Are Similar in Resting and Activated B Cells*).

DISCs Have Sequence Characteristics Consistent with H-DNA. Many repetitive sequences contain motifs with propensity to form non–B-form structures that contain intramolecular ssDNA (Fig. 2*A*) (9, 41–44). We therefore speculated that DISCs may correspond to locations where non-B DNA forms readily. Based on the biochemical and biophysical properties of non-B DNA and in vitro experiments using plasmids and oligonucleotides, algorithms have been developed to identify sequences that have potential for non-B DNA isomerization (45–47). When annotated non-B DNA motifs in the mouse genome (34) were compared with S1-seq signal for *Spo11*$^{-/-}$ mice, H-DNA motifs harbored >50% of total S1-seq reads, much more than for other motifs (Fig. 2*B*). Moreover, >99% of 144,478 DISCs were located on putative H-DNA–forming sequences (annotated H-DNA motifs plus other pyrimidine mirror repeats) (Fig. 2*C*).

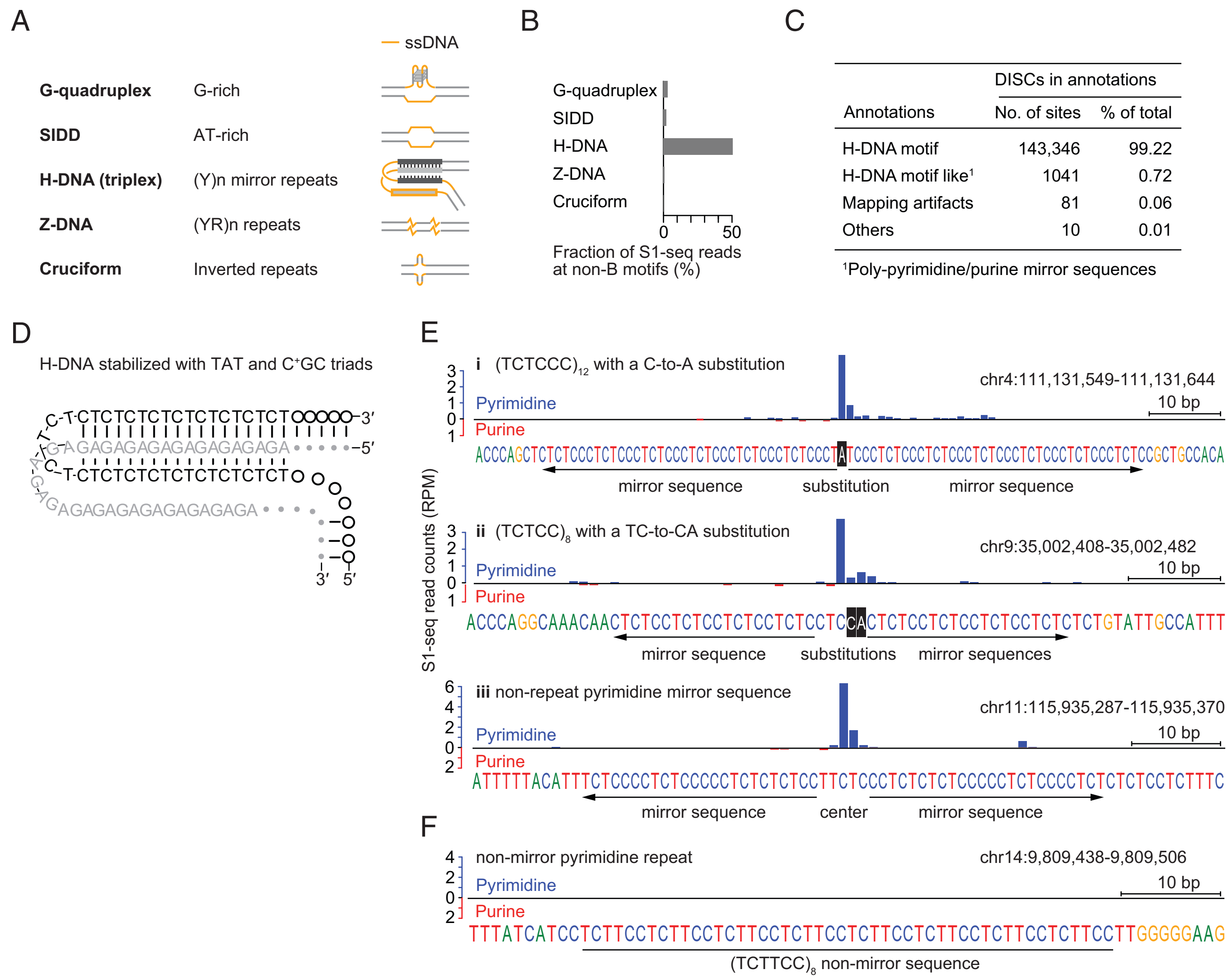
Stable intramolecular triplex is favored by formation of Hoogsteen triads (*SI Appendix*, Fig. S2*A*), continuous runs of which can be formed by pyrimidine mirror sequences (Fig. 2*D*) (9). Disrupting pyrimidine mirror repeats by base substitutions affects the H-DNA–forming potential of the repeats in plasmids, whereas substitutions located centrally in between mirror repeats do not (48). Consistent with these properties, S1-seq reads were still enriched at polypyrimidine runs that had a centrally located substitution(s) that did not disrupt the mirror repeats (Fig. 2 *E*, *i* and *ii*). S1-seq reads were also enriched at nonrepeat pyrimidine mirror sequences (Fig. 2 *E*, *iii*). However, nonmirror pyrimidine repeats showed little or no enrichment (Fig. 2*F* and *SI Appendix*, Fig. S2*B*). These results suggest that H-DNA–forming potential per se predicts S1-seq enrichment.

S1-seq Patterns May Reflect Properties of Specific H-DNA Isomers. We sought to discern how H-DNA structures might give rise to the observed S1-seq patterns, namely, the strand asymmetry and central position of the major cluster of reads at pyrimidine mirror sequences (Fig. 1*H*). A pyrimidine mirror sequence can fold into four isomers (H-y5, H-y3, H-r5, and H-r3) named after the origin of the third strand in the triplex: whether it comes from the pyrimidine or purine strand and whether from the 5′ or 3′ half of that strand (Fig. 3*A*) (9).

Complete digestion of central loop ssDNA with nuclease S1 and dissociation of the triplex should leave (among other things discussed below) a duplex DNA end at the border between the central loop and the original triplex (Fig. 3*A*). For isomers H-y5 and H-r3, this duplex DNA end should allow adaptor ligation to the right half of the mirror sequence, resulting in a centrally located pyrimidine-strand S1-seq read (Fig. 3 *A*, *Right*, blue arrow). This matches the observed pattern in S1-seq data (Fig. 1*H*); thus, H-y5 and H-r3 can plausibly account for the major central pyrimidine-strand S1-seq signal. In contrast, digestion of isomers H-y3 and H-r5 should allow adaptor ligation to the left half of the mirror sequence, resulting in a central purine-strand read (Fig. 3 *A*, *Left*, red arrow). This was not observed at high frequency, suggesting that these isomers are unlikely to be sources of the S1-seq signal.

Polypyrimidine mirror sequences can also make non–B-form structures other than H-DNA, namely, nodule DNA and slipped DNA, which are tandem repeats of triplexes or of ssDNA loops, respectively (Fig. 3*B*) (49, 50). These structures are symmetric, unlike H-DNA, so they should produce equivalent numbers of S1-seq reads from both strands. Moreover, such reads would not be centrally located within the mirror sequence (Fig. 3*B*). Therefore, neither structure is a good candidate to explain the central S1-seq signal.

DISCs show strong asymmetry between the strands for total S1-seq read count in and around pyrimidine mirror sequences (Fig. 1*H*). If S1 cleaves both strands of the H-DNA junction region, it should leave a duplex DNA end that flanks the mirror



	DISCs in annotations	
Annotations	No. of sites	% of total
H-DNA motif	143,346	99.22
H-DNA motif like[1]	1041	0.72
Mapping artifacts	81	0.06
Others	10	0.01

[1]Poly-pyrimidine/purine mirror sequences

Fig. 2. DISCs correspond to sites with characteristics consistent with H-DNA. (*A*) Schematic examples of non–B-form DNA structures containing ssDNA (yellow lines): G-quadruplex, stress-induced duplex destabilized site (SIDD), H-DNA, Z-DNA, and cruciform. (*B*) Fraction of S1-seq reads from *Spo11*$^{-/-}$ mice at non–B-form DNA motifs. Motif annotations are from ref. 34. (*C*) Overlap of DISCs with H-DNA motifs. DISCs were called as peaks with >0.65 RPM per 50 bp in *Spo11*$^{-/-}$ S1-seq data and compared with annotated H-DNA motifs (34). "H-DNA motif like" refers to polypyrimidine mirror repeats that were not annotated as H-DNA motifs. "Mapping artifacts" are DISCs next to genome assembly gaps which therefore are likely to be within poorly assembled repetitive DNA, creating artificial spikes in the S1-seq maps. "Others" are DISCs that could not be classified into these categories. (*D*) Schematic of an example of H-DNA showing base sequence. H-DNA is stable if the constituent triads are TAT or C$^+$GC in H-y isomers (shown) or are TAA or CGG in H-r isomers. (*E* and *F*) Examples of S1-seq read patterns around different types of polypyrimidine sequences. (*i* and *ii*) Imperfect polypyrimidine repeats in which the interruptions are centrally located. (*iii*) A mirror-symmetric polypyrimidine sequence that is not a direct repeat. (*F*) A polypyrimidine repeat that is not mirror symmetric.

sequence and that can be ligated to an adaptor, giving rise to a sequencing read pointed away from the mirror sequence (Fig. 3*C*). This would yield a purine-strand read on the left side for H-y5 and H-r3 (the inferred major source of DISC signal) or a pyrimidine-strand read on the right for H-y3 and H-r5 (Fig. 3*C*). Such reads consistent with H-y5 and H-r3 were indeed observed (Fig. 1*H*). However, if S1 cleavage of junctions is efficient, these reads should be comparable in number to the reads from cleavage at the central loop, but we instead observed a much smaller number of these candidate junction reads (Fig. 1 *E*, *G*, and *H*).

A straightforward way to account for this difference is if S1 does not always fully digest all of the ssDNA. For example, if S1 cleaves the loop and orphan strand but not the junction strand, this would leave a long 3′ overhang for H-y5 that might block ligation if it is inefficiently removed during the subsequent polishing step with T4 DNA polymerase (*SI Appendix*, Fig. S3*A*). If the junction strand is less efficiently digested than the other strands, this scenario would represent a majority of S1-digested H-DNA molecules, which might then account for the relative deficit of junction reads as compared to central reads. Additionally, if H-DNA were sometimes digested on the orphan and junction strands but leaving the central loop intact, this would be expected to leave a long 5′ overhang for H-y5 that, after fill-in by T4 DNA polymerase, would be predicted to yield a pyrimidine-strand junction read mapping just to the left of the mirror repeats (Fig. 3*D*). Pyrimidine-strand reads matching this expectation are apparent on the left side of C(TC)$_{20}$ repeats (Fig. 1 *G* and *H*). Depending on which strands end up being cleaved by S1 and where those cleavages occur, additional detailed features of the DISK S1-seq patterns can be readily explained, including the periodically spaced purine-strand reads inside the left half of mirror sequences (*SI Appendix*, Fig. S3*B*).

To further test whether DISCs might reflect specific H-DNA structures, we compared the location and strength of

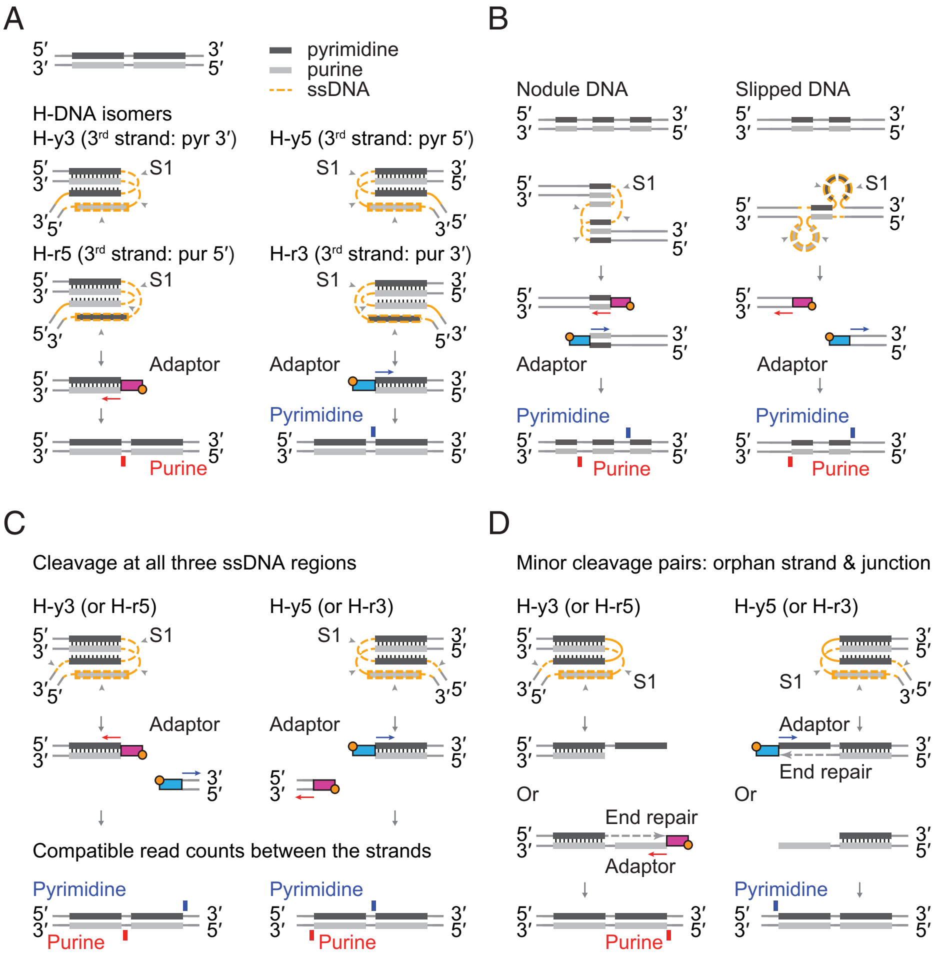

A
5′ 3′
3′ 5′
pyrimidine
purine
ssDNA
H-DNA isomers
H-y3 (3rd strand: pyr 3′)
H-y5 (3rd strand: pyr 5′)
S1
H-r5 (3rd strand: pur 5′)
H-r3 (3rd strand: pur 3′)
Adaptor
Pyrimidine
Purine
B
Nodule DNA
Slipped DNA
C
Cleavage at all three ssDNA regions
H-y3 (or H-r5)
H-y5 (or H-r3)
Compatible read counts between the strands
D
Minor cleavage pairs: orphan strand & junction
End repair
Or

Fig. 3. Models to illustrate how S1-seq strand asymmetry can be explained as a readout of H-DNA structure. In all panels, gray arrowheads indicate which ssDNA segments are digested with nuclease S1, and the adaptors are color coded to indicate whether the resulting sequencing read will map to the pyrimidine (blue) or purine strand (red). At the bottom of each panel, the schematic indicates the expected mapping position and strand for the S1-seq read. (*A*) H-DNA can exist as any of four isomers. Each isomer is named after the third strand in the duplex; e.g., if the third strand is from the 5′ half of the pyrimidine strand, the isomer is called H-y5. For H-y5 and H-r3 (*Right*), S1 digestion to yield a duplex DNA end at the border between the central loop and the original triplex should allow adaptor ligation to the right half of the mirror sequence, resulting in a centrally located pyrimidine-strand S1-seq read (blue). In contrast, digestion of isomers H-y3 and H-r5 should allow adaptor ligation to the left half of the mirror sequence, resulting in a central purine-strand read (red). (*B*) Other possible non–B-form structures, nodule DNA and slipped DNA, were not suitable for explaining the position and strand asymmetry of S1-seq reads. (*C*) Expected S1-seq read positions if ssDNA of H-DNA is fully digested. For H-y5 and H-r3 (*Right*), digestion of both strands at the junction should yield a flanking purine-strand read in addition to the central pyrimidine-strand read detailed in *A*. For H-y3 and H-r5 (*Left*), the junction read should map to the pyrimidine strand. If junction digestion is efficient, then comparable read counts are expected at the junction and central positions. (*D*) Expected S1-seq read positions if H-DNA is only partially digested by nuclease S1 on the orphan strand and junction. Partial digestion at the indicated positions on H-y5 (*Right*) would yield a 5′ overhang that could be filled in by T4 DNA polymerase and ligated to a sequencing adaptor, yielding an S1-seq read that maps to the pyrimidine strand just to the left of the mirror repeat. For H-y3, this scenario predicts a purine-strand read to the right of the mirror repeat, while H-r5 and H-r3 would yield 3′ overhangs. If these are inefficiently polished, they would not yield a sequencing read (as shown); if polished and ligated, they would yield a central read indistinguishable from S1 having cleaved the central loop (as in *A*).

the S1-seq signal with the previously determined sensitivity of an H-y5–forming plasmid to chemical probes $KMnO_4$, which detects unpaired T and C, and acid depurination, which detects unpaired G and A. Glover et al. studied the chemical reactivity of the H-y5 isomer formed on $(TC)_{17}$ (51). On the pyrimidine strand, chemical reactivity was highest in the central loop region, followed by flanking sequence to the left, then within the 5′ (left) half of the repeat (*SI Appendix*, Fig. S3*C*). On the purine strand, chemical reactivity was highest in the central loop region and the orphan strand (the 3′ [left] half of the repeat), then the 5′ (right) half of the repeat, then downstream to the right. The greater reactivity of the central loop than the junction sequence is consistent with our hypothesis that the central loop might also be more easily digested by nuclease S1. The modest chemical reactivity of both the purine strand and the third strand within the triplex suggests that even though fully incorporated in the triplex in the H-DNA structure model, they may be partially unpaired in solution and thus may also be cleaved by nuclease S1. Taken together, the S1-seq results appear largely congruent with expectation based on chemical probes of single-strandedness in H-y5 H-DNA.

Nevertheless, we noticed rare examples that might represent (mixtures of) other H-DNA isomers. Loop sequence plays a crucial role for isomerization of intramolecular DNA triplexes in supercoiled plasmids (52, 53). *SI Appendix*, Fig. S4*A* shows an example of a pyrimidine mirror repeat with a 10-bp interruption between the repeats. The S1-seq signal is markedly different from the pattern at more contiguous pyrimidine mirror repeats, with central signal enriched on both strands. This pattern could be consistent with alternative H-DNA isomers (i.e., H-y3 or H-r5 in addition to the more prevalent H-y5 or H-r3).

S1-seq Patterns at Pyrimidine Mirror Repeats with Different Sequence Compositions. H-DNA studies began with the discovery that some plasmids showed reactivity to nuclease S1 (54). Subsequently, H-DNA structural features have been extensively studied using plasmids in which various potential H-DNA–forming sequences were individually cloned and characterized (55). However, the number of sequences analyzed per study was of necessity limited, and whether results with a given set of sequences could be extrapolated to untested sequences remains unclear. S1-seq of mouse genomic DNA allows us to systematically analyze many thousands of pyrimidine mirror repeats over the mouse genome simultaneously.

First, we consider the effect of repeat sequence on S1-seq signal. Because S1-seq read counts varied between different pyrimidine mirror sequences (*SI Appendix*, Fig. S4*B*), we hypothesized that S1-seq signal reflects the frequency and stability of triplex structure formation. To test this, we compared the S1-seq signal intensity with the previously reported ease of H-DNA formation in plasmids. Plasmids with $(TCCTC)_n$ require greater superhelical stress per unit length to form H-DNA than with the repeat $(TC)_n$ (56). Congruently, we found that the S1-seq signal intensity was greater for TC repeats than for TCCTC repeats of the same length [compare $C(TC)_{24}$ with $TC(TCCTC)_9T$ in *SI Appendix*, Fig. S2*B*].

Moreover, detailed spatial patterns of S1-seq reads differed between different repeat sequences. Specifically, each type of repeat showed a periodically spaced S1-seq signal that was highly stereotyped for a given repeat sequence but that differed between repeats of different sequence (Fig. 1*H* and *SI Appendix*, Fig. S4*C*). Analogously, Collier and Wells found that plasmid-borne $(TCCTC)_{12}$ gave a periodic pattern of chemical sensitivity with peaks every five bases (56). A simple single-base preference for the chemical probes or nuclease S1 alone would not explain why different repeat sequences give peaks at different positions, so we surmise that local differences in triplex structure modulate nuclease sensitivity and chemical reactivity.

Variations in S1-seq Patterns According to Pyrimidine Mirror Repeat Length. Next, we consider the effect of repeat length on S1-seq signal. How the lengths of mirror repeats shape H-DNA three-dimensional structures remains largely uncharacterized. Previous plasmid studies suggested that long polypyrimidine stretches may form large H-DNA (57) or nodule DNA (two tandem H-DNAs) (49). To address this issue, we focused on TC repeats of various lengths because these repeats are abundant in the mouse genome and are highly enriched for S1-seq reads.

Fig. 4*A* shows a heat map of strand-specific S1-seq reads for $C(TC)_n$ repeats with n ranging from 16 to 38 in steps of 2, centered on the repeat midpoints. The read distributions were strongly stereotyped for copies of the same length but showed progressive differences that tracked with repeat length. Specifically, the strong cluster of pyrimidine-strand reads inferred to be from cleavage at the central loop grew wider as TC repeat length increased, and the weaker clusters of pyrimidine- and purine-strand reads inferred to represent junction cleavage moved progressively leftward.

We considered two models to explain the expansion of the central S1-seq signal as repeat length increased, taking into account that $C(TC)_{16}$ is the minimum needed to form detectable H-DNA (58). In one model (the sliding model, shown in Fig. 4 *B*, *Left*), the triplex segment can be positioned anywhere within a repeat that is longer than the minimum (i.e., $n \geq 16$). As a result, the central loop can occupy any position within the repeat except within $n = 8$ (16 bp) from the ends, and the central S1-seq signal becomes wider at longer repeats because it is the superposition of a population of alternative H-DNA structures that can be differentiated from each other by sliding the triplex across the repeat. An alternative model (the end-fixed model, shown in Fig. 4 *B*, *Right*) envisions that the triplex position is fixed at the left and right ends of the repeat. In this model, the longer the repeat is, the longer the central loop will be on average, and the wider the distribution of central loop boundary positions will be. This model thus explains the spread of the central S1-seq signal as the superposition of a population of H-DNA structures that have the triplex position fixed at the ends but variable positions for the boundary of the central loop. In both models, the triple-stranded region can vary in length.

To determine which model is better suited to explain the overall signal pattern, we focused on the clusters of junction reads immediately to the left of the repeat on both strands. The sliding model predicts that the junction position will be variable as the repeat length increases, spreading from the left end of the repeat into the repeat itself. In contrast, the end-fixed model explicitly posits a uniform junction position just outside the left end of the repeat, regardless of the length of the repeat. When the heat maps were centered on the left end of the polypyrimidine tracts, the junction S1-seq signal remained confined to one position irrespective of the number of repeats (Fig. 4*C*), consistent with the end-fixed model.

Various Intramolecular Triplex Structures Form on Long TC Repeats. For longer TC repeats ($n \geq 24$), the central S1-seq signal on the pyrimidine strand showed pronounced substructure (Fig. 4*A*), with multiple peaks when repeats of the same length were averaged (Fig. 4*D*). This substructure cannot be explained by a simple end-fixed model in which the triplex length is also fixed and the central loop region is entirely ssDNA. We therefore hypothesized instead that the complex S1-seq read distribution reflects an average over a population of alternative structures that differ with respect to which portions of the central loop are single stranded.

To test this idea, we examined the effects of repeat-disrupting substitutions. As noted in *DISCs Have Sequence Characteristics Consistent with H-DNA*, substitution is tolerated within the central loop (Fig. 2*E*), but a substitution within the triplex-forming region disfavors H-DNA formation because the substitution prevents Hoogsteen triad formation (48, 59). We reasoned that a repeat-disrupting substitution would limit the variety of alternative structures that might be formed by a long TC repeat because the triple-stranded region would be constrained not to overlap the substitution. We therefore searched for imperfect $C(TC)_n$ repeats in which one cytosine position was instead a different base and then examined how S1-seq read patterns changed depending on the position of the substitution.

Heat maps for imperfect $C(TC)_n$ repeats are shown in Fig. 4 *E*–*G* for $n = 20$, 26, and 30, respectively. The leftmost C in $C(TC)_n$ is indicated as C0, and each cytosine is named from C0 to Cn in order from left to right. As predicted, each substitution position was associated with a marked and stereotyped difference in the S1-seq read distribution. Substitutions near the center of the repeat [e.g., C10 to C21 for $C(TC)_{30}$] gave pronounced clusters of pyrimidine-strand S1-seq reads immediately around the substitution, as expected if the substitution is constrained to be within the central loop and is often single-stranded.

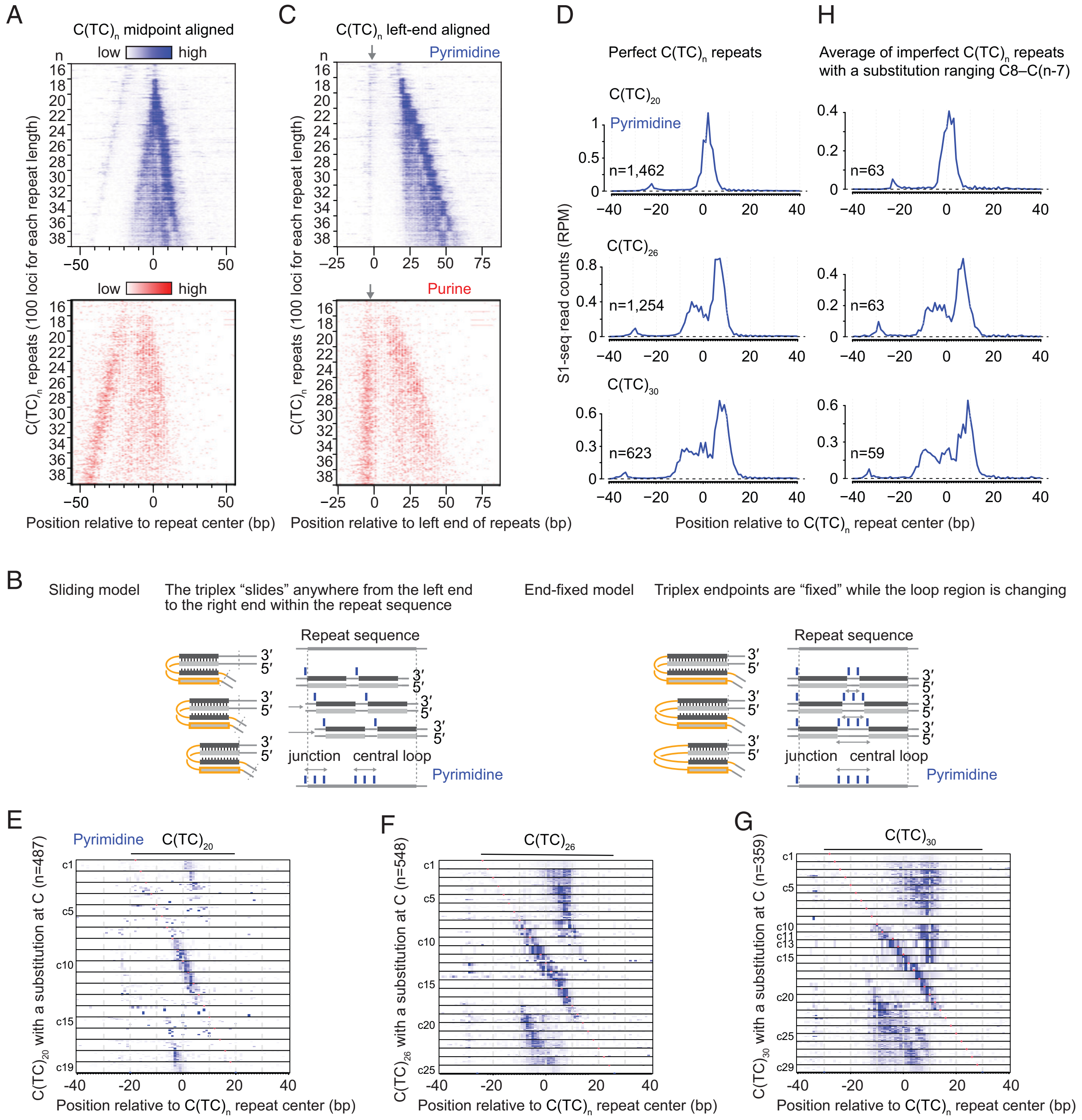


Fig. 4. Variation in S1-seq read patterns with differences in TC repeat length and composition. (*A*) Strand-specific heat maps of S1-seq signal for C(TC)$_n$ repeats with n ranging from 16 to 38 in steps of 2, centered on the repeat midpoints. Color gradients for the pyrimidine strand (blue) and purine strand (red) are scaled differently to illustrate each strand's spatial patterns. For each value of n, 100 loci of that length were randomly selected for display. (*B*) Alternative models to explain the spread of pyrimidine-strand central signal as TC repeat length increases. See *Variations in S1-seq Patterns According to Pyrimidine Mirror Repeat Length* for details. The triplex length is shown as constant for simplicity, but it can vary in either model. (*C*) Fixed position of junction reads on both strands irrespective of TC repeat length. The heat maps show strand-specific S1-seq signal for the same C(TC)$_n$ repeats as in *A* but lined up along the start of the repeat. Gray arrows indicate the position of the presumed junction reads. (*D*) Averaged pyrimidine-strand S1-seq signal for perfect C(TC)$_n$ repeats of the indicated lengths. (*E*–*G*) S1-seq patterns on imperfect C(TC)$_n$ repeats differ in stereotyped ways depending on the position of the imperfection and the length of the repeat. Heat maps show pyrimidine-strand S1-seq maps for imperfect C(TC)$_{20}$ (*E*), C(TC)$_{26}$ (*F*), and C(TC)$_{30}$ (*G*). Each repeat analyzed has a single cytosine substitution, and the substitution positions are grouped and ordered with the leftmost substitution (C1) at the top. (*H*) Averaged pyrimidine-strand S1-seq signal for imperfect C(TC)$_n$ repeats of the indicated lengths. For each graph, repeats with a single cytosine substitution ranging from C8 to C(n-7) were averaged.

If the complex pattern for the central S1-seq signal on long perfect TC repeats reflects a population of alternative structures with ssDNA occupying different positions, we reasoned that we could mimic this more complex population by combining the simpler populations observed for the imperfect TC repeats whose central loops are more constrained. Indeed, when we created plots averaging the signal for imperfect repeats with central substitutions, these recapitulated well the average plots for perfect repeats for the pyrimidine strand (Fig. 4*H*) and the purine strand (*SI Appendix*, Fig. S5 *A* and *B*).

S1-seq Patterns at H-DNA Motifs Are Similar in Resting and Activated B Cells. To ask whether DISCs might reflect H-DNA that was already present in vivo, we measured S1-seq patterns around H-DNA motifs in DNA isolated from cultured primary mouse splenic B cells, comparing transcriptionally quiescent resting cells to transcriptionally active cells that had been stimulated by treatment ex vivo with lipopolysaccharide and interleukin 4 (LPS + IL4) (60). Previous studies using ssDNA-seq—which uses $KMnO_4$ and nuclease S1 treatment to introduce DSBs at locations enriched for unpaired bases in vivo—detected a broad zone of sequencing read enrichment around H-DNA motifs in activated B cells, compared with a narrower band of relative depletion in resting cells (Fig. 5*A*) (34). We therefore reasoned that S1-seq patterns would also be different between the two cell states if S1-seq were detecting transcription-promoted H-DNA that had formed in vivo. However, average S1-seq signal intensity at H-DNA motfis was highly similar between activated and resting B cells (Fig. 5*A*), and spatial patterns and read density were also similar when considering just $C(TC)_{20}$ repeats (Fig. 5*B* and *SI Appendix*, Fig. S6 *A* and *B*). These data thus did not provide support for the idea that S1-seq captures intramolecular triplex structures that were formed in vivo. One possibility is that most or all of the triplexes detected by S1-seq form after DNA isolation. Alternatively, H-DNA detected by S1-seq may form in vivo but independently of transcriptional status. Importantly, these findings speak only to what S1-seq detects; they do not exclude that H-DNA does indeed form in vivo.

If the H-DNA detected by S1-seq was formed ex vivo, a couple of observations suggest that having a pyrimidine mirror sequence may not be sufficient to yield detectable signal. First, different TC repeats of the same length yielded widely different amounts of S1-seq signal (*SI Appendix*, Fig. S2*B*). Second, the single TC repeat longer than $C(TC)_{20}$ in the yeast genome did not yield characteristic H-DNA S1-seq patterns (*SI Appendix*, Fig. S6*C*).

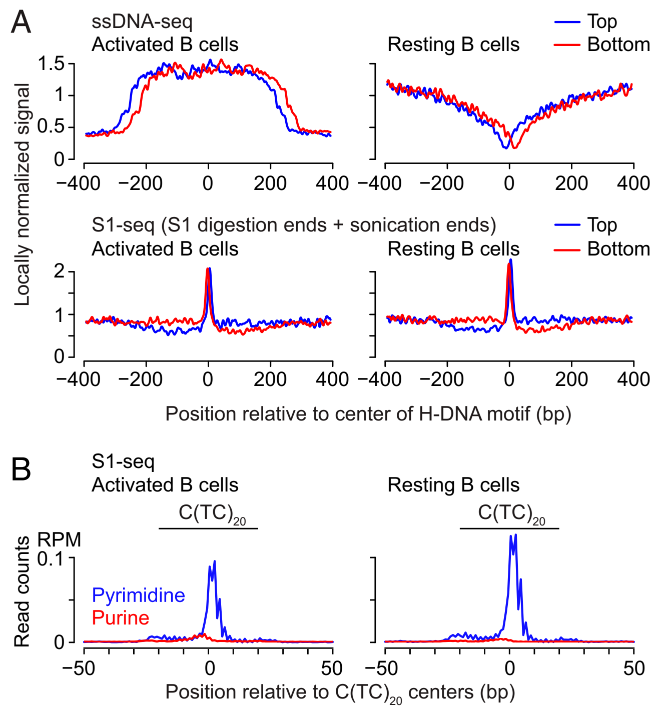


Fig. 5. S1-seq patterns at H-DNA motifs in activated vs. resting splenic B cells. (*A*) Locally normalized average plots of ssDNA-seq (34) and S1-seq for activated and resting B cells, centered on H-DNA motifs that were previously annotated (34). (*B*) Average plot of S1-seq for activated and resting B cells, centered on $C(TC)_{20}$ loci.

Phylogenetic Aspects of TC Repeats. Noticing that mice have many more TC repeats than humans prompted us to investigate the phylogenetic distribution of TC repeats. $C(TC)_{20}$ is abundant (>2,500 copies) in the genome assemblies of mouse, rat, and Chinese hamster but is much less common in more distantly related species including kangaroo rat, naked mole-rat, ground squirrel, and pika (Fig. 6*A*).

For many mouse $C(TC)_{20}$ repeats, the sequence immediately 3′ on the pyrimidine strand is enriched for additional degenerate TC repeats, but a substantial subset [648 out of 2,878 total $C(TC)_{20}$ repeats] instead has an AC repeat in either forward or reverse orientation (i.e., either a TG or AC repeat) (Fig. 6*B*). The same is true for rat and Chinese hamster (Fig. 6*C*). Moreover, more than 60% of mouse $C(TC)_{20}$ repeats are conserved in rat and Chinese hamster (Fig. 6 *D*, *Top*), and, conversely, more than 60% of such repeats in rat and Chinese hamster are conserved in mouse (Fig. 6 *D*, *Bottom*). Some $C(TC)_{20}$ repeats in kangaroo rat, naked mole-rat, and squirrel were also followed by AC repeats, but downstream TG repeats were rare (Fig. 6*C*), and $C(TC)_{20}$ repeat positions were less well conserved with mice (Fig. 6*D*). In the human genome, $C(TC)_{20}$ repeats were often followed by AT or, less often, TG repeats (Fig. 6*C*). These findings indicate that TC repeats amplified and spread throughout the genome at or before the last common ancestor of mice, rats, and Chinese hamster. Moreover, at least a subset of TC repeats were in the context of a more complex repeat structure at the time of amplification.

Discussion

In this study, we uncovered features of apparent triplex structures within the mouse genome by using S1-seq signal as a footprint of non–B-form DNA structure. DISCs—S1-seq hot spots that appear to be unrelated to the presence of DSBs—were clearly correlated with the H-DNA–forming potential of their DNA sequences. Moreover, local S1-seq distributions were consistent with H-y5 and/or H-r3 isomers of H-DNA, including the strong central pyrimidine-strand signal and the strong strand asymmetry of the sequencing read count.

An open question is whether the triplex detected by S1-seq was already present in vivo or instead it formed ex vivo. A plausible argument can be made that most, if not all, of the triplexes that we detected were formed ex vivo because the removal of histones and other proteins from the chromatin while the DNA was constrained in agarose might have provided sufficient negative superhelicity to favor triplex formation. Moreover, the digestion with nuclease S1 was conducted at acidic pH in the presence of divalent cations, which both favor H-DNA formation (discussed further below). An independent study using a variant of S1-seq recently reported sequencing signal enriched at polypyrimidine•polypurine mirror repeats, which the authors interpreted as arising from H-DNA formed in vivo during DNA replication (61). We sought evidence in support of triplex formation in vivo by comparing S1-seq patterns on DNA from resting vs. activated B cells. The lack of a clear difference between the cell types indicates either that most of the DISC signal we detected was formed ex vivo or that H-DNA formation in vivo is independent of transcriptional status.

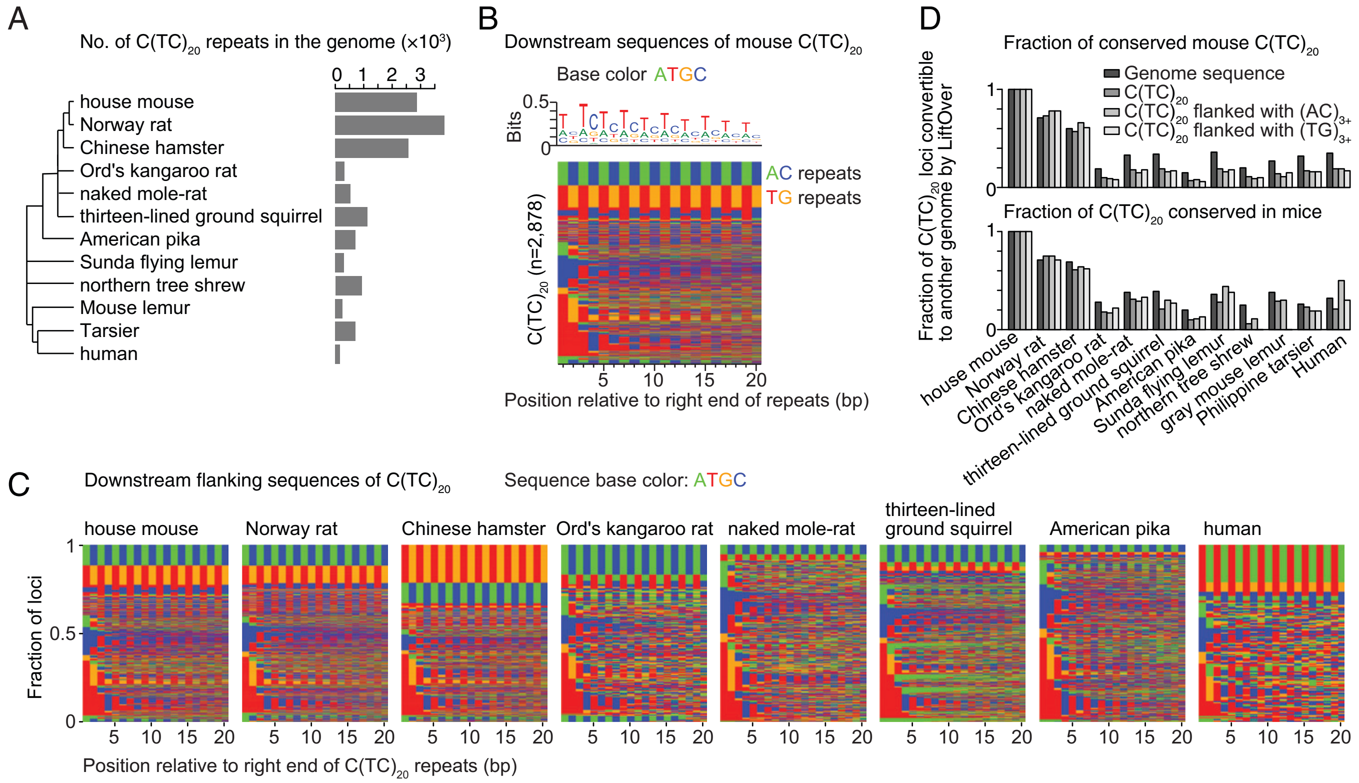


Fig. 6. Amplification of TC repeat copy number in rodents. (*A*) $C(TC)_{20}$ repeat copy number in various species: house mouse (C57BL/6J strain of *Mus musculus*), Norway rat (*Rattus norvegicus*), Chinese hamster (*Cricetulus griseus*), Ord's kangaroo rat (*Dipodomys ordii*), naked mole-rat (*Heterocephalus glaber*), thirteen-lined ground squirrel (*Spermophilus tridecemlineatus*), American pika (*Ochotona princeps*), Sunda flying lemur (*Galeopterus variegatus*), northern tree shrew (*Tupaia belangeri*), gray mouse lemur (*Microcebus murinus*), Philippine tarsier (*Tarsius syrichta*), and human (*Homo sapiens*). Tree topology is from the UCSC genome browser; branch lengths are not to scale. (*B*) Sequence context of mouse $C(TC)_{20}$ repeats. The clustered color map shows the 20 nucleotides 3′ of $C(TC)_{20}$ repeats in mice. Green, blue, red, and orange indicate adenine, cytosine, thymine, and guanine, respectively. (*C*) Sequence context of $C(TC)_{20}$ repeats in other species. The clustered color maps are presented as in *B*. (*D*) Conservation of $C(TC)_{20}$ sequences. (*Top*) The fraction of mouse $C(TC)_{20}$ sequences (all copies, or subsets that are followed by TG or AC repeats, as indicated) that are conserved in each of the indicated species. (*Bottom*) The fraction of such $C(TC)_{20}$ sequences from the indicated species that are conserved in mouse. Black bars show overall genomic sequence conservation as a point of comparison.

We favor the idea that most of the H-DNA that we detected by S1-seq was formed ex vivo. Importantly, however, we do not exclude the possibility that at least some triplex was already present in vivo. Moreover, even if all of the triplex detected by S1-seq had formed ex vivo, it would not exclude the possibility that triplex does indeed form in vivo. Nevertheless, our findings establish that S1-seq can be useful as a tool to probe H-DNA potential genome-wide at high resolution.

The S1-seq signal is consistent with both H-y5 and H-r3 isomers, but the following considerations led us to infer that the triplex detected by S1-seq is probably H-y5. For formation of H-y isomers, acidic pH and longer length are favorable (9, 56). Moreover, the H-y5 isomer is observed at lower superhelical density than H-y3 and can even occur on linearized plasmids (62). Also, the presence of divalent cations (e.g., Ca^{2+}, Zn^{2+}, Mg^{2+}, and Mn^{2+}) makes the H-y5 isomer preferable (63). Relevant to these points, our nuclease S1 digestion buffer is acidic (pH 4.5) and contains 4.5 mM Zn^{2+}.

We note that the triplex-forming sequences detected by S1-seq in this study are longer than most of the sequences that have been examined in plasmids in previous studies. For TC repeats longer than $(TC)_{15}$, one negative superhelical turn in plasmid DNA is relieved for every 11 nucleotides of $(TC)_n$ repeat that can be converted from duplex to the H-DNA conformation (58). TC repeats shorter than $(TC)_{13}$ relieve more superhelical turns to form H-DNA, suggesting that the longer repeats [longer than $(TC)_{15}$] can form H-DNA at lower superhelical density. We found that S1-seq signal was enriched at TC repeats longer than $C(TC)_{16}$, consistent with triplex formation that requires only a low superhelical density. The H-y5 structure is favored under conditions of very low or no topological stress (explained by higher stacking energy in the H-y5 conformer), whereas high topological stress generated only the H-y3 conformer (which relaxes more supercoil writhe) (64). These considerations lend further support to the conclusion that S1-seq is detecting primarily the H-y5 conformation and also reinforce the plausibility of H-DNA forming on the agarose-embedded linear genomic DNA under the conditions of S1 digestion.

S1-seq (this study) and ssDNA-seq (34) both detect ssDNA but gave different results at H-DNA motifs. First, S1-seq signal was enriched at most of the long H-DNA motifs, but ssDNA-seq was enriched at only a subset of the motifs and showed only a weak enrichment at TC repeats. Second, ssDNA-seq showed marked differences between resting and activated B cells, but S1-seq did not. Third, DSB-independent S1-seq signal was almost exclusively at H-DNA motifs, but ssDNA-seq signal was observed at H-DNA as well as other non-B motifs (34). These differences presumably reflect the different methodologies for probing ssDNA. In particular, ssDNA-seq uses nuclease S1 to detect sites where ssDNA was previously modified in vivo with permanganate, whereas S1-seq relies on direct digestion of unmodified DNA with nuclease S1.

Our findings show that S1-seq can be used to study detailed triplex structures on mammalian genomic DNA. One strength is that S1-seq permits examination of a large number of intramolecular-triplex–forming sequences on genomic DNA at once. Thus, this study may serve as a basis for investigating the function of triplex structures on genomic DNA.

Methods

Mice. Experiments conformed to the US Office of Laboratory Animal Welfare regulations and were approved by the Memorial Sloan Kettering Cancer Center Institutional Animal Care and Use Committee. Mice were maintained on regular rodent chow with continuous access to food and water until euthanasia by CO_2 asphyxiation prior to tissue harvest. C57BL/6J males were obtained from the Jackson Laboratory.

S1-seq.

B cell preparation. Splenic B cells were obtained as described previously (60). Briefly, resting naive mouse B cells were isolated from splenocytes with anti-CD43 MicroBeads (Miltenyi Biotec) by negative selection and activated for 72 h in the presence of LPS (50 μg/mL final concentration, *Escherichia coli* 0111:B4; Sigma-Aldrich) plus IL4 (2.5 ng/mL final concentration; Sigma-Aldrich). Apoptotic cells were removed with a Dead Cell Removal Kit (Miltenyi) followed by Ficoll gradient with >90% live-cell purity.

DNA extraction in plugs. Cells were embedded to protect DNA from shearing, and DNA was liberated by treatment with SDS and proteinase K as described (7). In-plug overhang removal with nuclease S1 and adaptor ligation were performed as described (35, 36). After ligation to biotinylated adaptors, DNA was purified from the agarose, sheared by sonication, purified with streptavidin, ligated to second-end adaptors, amplified, and sequenced as described (7, 35, 36).

Mapping and preprocessing. Sequence reads were mapped onto the mouse reference genome (mm10) by bowtie2 version 2.2.1 (65) with the argument –X 1000. Uniquely and properly mapped reads were counted, at which a nucleotide next to biotinylated adaptor DNA was mapped. Mapping statistics are in *SI Appendix*, Table S1.

Sequence Motif Search. Genome sequences were searched using dreg (European Molecular Biology Open Software Suite [EMBOSS] version 6.6.0.0) (66) with default parameters. Repeat length was defined as the longest repeating subsequence with no mismatches, insertion, or deletions; i.e., no $(TC)_{20}$ repeats were annotated within a $(TC)_{21}$ repeat. Genome versions used are in *SI Appendix*, Table S2. Genomic coordinates of repeats shown in figures are in Dataset S1.

Quantification and Statistical Analyses. Statistical analyses were performed using R version 3.3.1 to 3.6.1 (www.r-project.org). Repeatmasker was downloaded from University of California, Santa Cruz (UCSC), genome browser on 12 January 2021.

Data Availability. Raw and processed sequencing S1-seq data were deposited in the Gene Expression Omnibus (GEO) (accession no. GSE197669) (67). We used mouse SPO11-oligo and yeast S1-seq data from GEO accession nos. GSE84689 and GSE85253, respectively; ExoT-seq and mouse S1-seq data were from GSE141850 (7, 8, 35). The yeast S1-seq data were mapped onto the yeast reference genome (sacCer2) by bowtie2 version 2.2.1 (65), and only uniquely mapped reads were counted. Non-B DNA annotation was obtained from ref. 34.

ACKNOWLEDGMENTS. We thank members of the S.K. and Maria Jasin laboratories for discussions. We thank the Memorial Sloan Kettering (MSK) Integrated Genomics Operation for sequencing. MSK core facilities are supported by National Cancer Institute Cancer Center Support Grant P30 CA08748. R.S. was supported by Grant U54 CA137788 from the MSK/City College of New York partnership; J.C. was supported by grants from the NIH (Grants R01 AI072194, R01 AI124186, U54 CA137788, and P30 CA008748), the Ludwig Center for Cancer Immunotherapy, MSK Functional Genomics, and the Geoffrey Beene Cancer Center. Work in the S.K. laboratory was supported by NIH Grant R35 GM118092 (to S.K.).

1. M. J. Behe, An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes. *Nucleic Acids Res.* **23**, 689–695 (1995).
2. P. Bucher, G. Yagil, Occurrence of oligopurine.oligopyrimidine tracts in eukaryotic and prokaryotic genes. *DNA Seq.* **1**, 157–172 (1991).
3. A. Bacolla, R. D. Wells, Non-B DNA conformations as determinants of mutagenesis and human disease. *Mol. Carcinog.* **48**, 273–285 (2009).
4. S. Kasinathan, S. Henikoff, Non-B-Form DNA is enriched at centromeres. *Mol. Biol. Evol.* **35**, 949–962 (2018).
5. J. Zhao, A. Bacolla, G. Wang, K. M. Vasquez, Non-B DNA structure-induced genetic instability and evolution. *Cell. Mol. Life Sci.* **67**, 43–62 (2010).
6. W. M. Guiblet *et al.*, Non-B DNA: A major contributor to small- and large-scale variation in nucleotide substitution frequencies across the genome. *Nucleic Acids Res.* **49**, 1497–1516 (2021).
7. S. Yamada *et al.*, Molecular structures and mechanisms of DNA break processing in mouse meiosis. *Genes Dev.* **34**, 806–818 (2020).
8. J. Lange *et al.*, The landscape of mouse meiotic double-strand break formation, processing, and repair. *Cell* **167**, 695–708.e16 (2016).
9. S. M. Mirkin, M. D. Frank-Kamenetskii, H-DNA and related structures. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 541–576 (1994).
10. G. P. Schroth, P. S. Ho, Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. *Nucleic Acids Res.* **23**, 1977–1983 (1995).
11. R. M. Clark, S. S. Bhaskar, M. Miyahara, G. L. Dalgliesh, S. I. Bidichandani, Expansion of GAA trinucleotide repeats in mammals. *Genomics* **87**, 57–67 (2006).
12. A. Bacolla *et al.*, Long homopurine*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. *Nucleic Acids Res.* **34**, 2663–2675 (2006).
13. M. Lexa, T. Martinek, M. Brazdova, "Uneven distribution of potential triplex sequences in the human genome. In silico study using the R/Bioconductor package triplex" in *Bioinformatics 2014: Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms*, O. Pastor *et al.*, Eds. (SciTePress, 2014), pp. 80–88.
14. J. J. Bissler, Triplex DNA and human disease. *Front. Biosci.* **12**, 4536–4546 (2007).
15. E. Kejnovsky, V. Tokan, M. Lexa, Transposable elements and G-quadruplexes. *Chromosome Res.* **23**, 615–623 (2015).
16. A. Bacolla, J. A. Tainer, K. M. Vasquez, D. N. Cooper, Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Res.* **44**, 5673–5688 (2016).
17. I. Georgakopoulos-Soares, S. Morganella, N. Jain, M. Hemberg, S. Nik-Zainal, Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res.* **28**, 1264–1271 (2018).
18. G. Tóth, Z. Gáspári, J. Jurka, Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Res.* **10**, 967–981 (2000).
19. L. Zhang *et al.*, Conservation of noncoding microsatellites in plants: Implication for gene regulation. *BMC Genomics* **7**, 323 (2006).
20. M. D. Frank-Kamenetskii, S. M. Mirkin, Triplex DNA structures. *Annu. Rev. Biochem.* **64**, 65–95 (1995).
21. M. M. Krasilnikova, S. M. Mirkin, Replication stalling at Friedreich's ataxia (GAA)n repeats in vivo. *Mol. Cell. Biol.* **24**, 2286–2295 (2004).
22. L. S. Son, A. Bacolla, R. D. Wells, Sticky DNA: In vivo formation in *E. coli* and in vitro association of long GAA*TTC tracts to generate two independent supercoiled domains. *J. Mol. Biol.* **360**, 267–284 (2006).
23. P. S. Sarkar, S. K. Brahmachari, Intramolecular triplex potential sequence within a gene down regulates its expression in vivo. *Nucleic Acids Res.* **20**, 5713–5718 (1992).
24. G. M. Samadashwily, A. Dayn, S. M. Mirkin, Suicidal nucleotide sequences for DNA polymerization. *EMBO J.* **12**, 4975–4983 (1993).
25. G. Wang, K. M. Vasquez, Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13448–13453 (2004).
26. B. P. Belotserkovskii *et al.*, A triplex-forming sequence from the human c-MYC promoter interferes with DNA transcription. *J. Biol. Chem.* **282**, 32433–32441 (2007).
27. G. Wang, K. M. Vasquez, Models for chromosomal replication-independent non-B DNA structure-induced genetic instability. *Mol. Carcinog.* **48**, 286–298 (2009).
28. G. Wang, S. Carbajal, J. Vijg, J. DiGiovanni, K. M. Vasquez, DNA structure-induced genomic instability in vivo. *J. Natl. Cancer Inst.* **100**, 1815–1817 (2008).
29. J. S. Lee, G. D. Burkholder, L. J. Latimer, B. L. Haug, R. P. Braun, A monoclonal antibody to triplex DNA binds to eucaryotic chromosomes. *Nucleic Acids Res.* **15**, 1047–1061 (1987).
30. Y. M. Agazie, J. S. Lee, G. D. Burkholder, Characterization of a new monoclonal antibody to triplex DNA and immunofluorescent staining of mammalian chromosomes. *J. Biol. Chem.* **269**, 7019–7023 (1994).
31. Y. M. Agazie, G. D. Burkholder, J. S. Lee, Triplex DNA in the nucleus: Direct binding of triplex-specific antibodies and their effect on transcription, replication and cell growth. *Biochem. J.* **316**, 461–466 (1996).
32. M. Ohno, T. Fukagawa, J. S. Lee, T. Ikemura, Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies. *Chromosoma* **111**, 201–213 (2002).
33. I. Lubitz, D. Zikich, A. Kotlyar, Specific high-affinity binding of thiazole orange to triplex and G-quadruplex DNA. *Biochemistry* **49**, 3567–3574 (2010).
34. F. Kouzine *et al.*, Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. *Cell Syst.* **4**, 344–356.e7 (2017).
35. E. P. Mimitou, S. Yamada, S. Keeney, A global view of meiotic double-strand break end resection. *Science* **355**, 40–45 (2017).
36. E. P. Mimitou, S. Keeney, S1-seq assay for mapping processed DNA ends. *Methods Enzymol.* **601**, 309–330 (2018).
37. N. Hunter, Meiotic recombination: The essence of heredity. *Cold Spring Harb. Perspect. Biol.* **7**, a016618 (2015).
38. I. Lam, S. Keeney, Mechanism and regulation of meiotic recombination initiation. *Cold Spring Harb. Perspect. Biol.* **7**, a016634 (2014).

39. P. Cejka, L. S. Symington, DNA end resection: Mechanism and control. *Annu. Rev. Genet.* **55**, 285–307 (2021).
40. A. Canela *et al.*, Topoisomerase II-induced chromosome breakage and translocation is determined by chromosome architecture and transcriptional activity. *Mol. Cell* **75**, 252–266.e8 (2019).
41. Y. Timsit, D. Moras, Cruciform structures and functions. *Q. Rev. Biophys.* **29**, 279–307 (1996).
42. A. Herbert, A. Rich, Left-handed Z-DNA: Structure and function. *Genetica* **106**, 37–47 (1999).
43. C. J. Benham, C. Bi, The analysis of stress-induced duplex destabilization in long genomic DNA sequences. *J. Comput. Biol.* **11**, 519–543 (2004).
44. S. Burge, G. N. Parkinson, P. Hazel, A. K. Todd, S. Neidle, Quadruplex DNA: Sequence, topology and structure. *Nucleic Acids Res.* **34**, 5402–5415 (2006).
45. O. Kikin, L. D'Antonio, P. S. Bagga, QGRS Mapper: A web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.* **34**, W676-82 (2006).
46. R. Z. Cer *et al.*, Non-B DB: A database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res.* **39**, D383–D391 (2011).
47. G. Wang, S. Gaddis, K. M. Vasquez, Methods to detect replication-dependent and replication-independent DNA structure-induced genetic instability. *Methods* **64**, 67–72 (2013).
48. S. M. Mirkin *et al.*, DNA H form requires a homopurine-homopyrimidine mirror repeat. *Nature* **330**, 495–497 (1987).
49. I. G. Panyutin, R. D. Wells, Nodule DNA in the (GA)37.(CT)37 insert in superhelical plasmids. *J. Biol. Chem.* **267**, 5495–5501 (1992).
50. R. R. Sinden, M. J. Pytlos-Sinden, V. N. Potaman, Slipped strand DNA structures. *Front. Biosci.* **12**, 4788–4799 (2007).
51. J. N. Glover, C. S. Farah, D. E. Pulleyblank, Structural characterization of separated H DNA conformers. *Biochemistry* **29**, 11110–11115 (1990).
52. M. Shimizu, K. Kubo, U. Matsumoto, H. Shindo, The loop sequence plays crucial roles for isomerization of intramolecular DNA triplexes in supercoiled plasmids. *J. Mol. Biol.* **235**, 185–197 (1994).
53. S. Kang, R. D. Wells, Central non-Pur.Pyr sequences in oligo(dG.dC) tracts and metal ions influence the formation of intramolecular DNA triplex isomers. *J. Biol. Chem.* **267**, 20887–20891 (1992).
54. H. Weintraub, A dominant role for DNA secondary structure in forming hypersensitive structures in chromatin. *Cell* **32**, 1191–1203 (1983).
55. R. D. Wells, D. A. Collier, J. C. Hanvey, M. Shimizu, F. Wohlrab, The chemistry and biology of unusual DNA structures adopted by oligopurine.oligopyrimidine sequences. *FASEB J.* **2**, 2939–2949 (1988).
56. D. A. Collier, R. D. Wells, Effect of length, supercoiling, and pH on intramolecular triplex formation. Multiple conformers at pur.pyr mirror repeats. *J. Biol. Chem.* **265**, 10652–10658 (1990).
57. Y. J. Han, P. de Lanerolle, Naturally extended CT. AG repeats increase H-DNA structures and promoter activity in the smooth muscle myosin light chain kinase gene. *Mol. Cell. Biol.* **28**, 863–872 (2008).
58. H. Htun, J. E. Dahlberg, Topology and formation of triple-stranded H-DNA. *Science* **243**, 1571–1576 (1989).
59. B. P. Belotserkovskii *et al.*, Formation of intramolecular triplex in homopurine-homopyrimidine mirror repeats with point substitutions. *Nucleic Acids Res.* **18**, 6621–6624 (1990).
60. W. T. Yewdell *et al.*, A hyper-IgM syndrome mutation in activation-induced cytidine deaminase disrupts G-quadruplex binding and genome-wide chromatin localization. *Immunity* **53**, 952–970.e11 (2020).
61. G. Matos-Rodrigues *et al.*, Linking dynamic DNA secondary structures to genome instability. bioRxiv [Preprint] (2022). https://www.biorxiv.org/content/10.1101/2022.01.19.476973v1 (Accessed 21 January 2022).
62. H. Htun, J. E. Dahlberg, Single strands, triple strands, and kinks in H-DNA. *Science* **241**, 1791–1796 (1988).
63. S. M. Kang, F. Wohlrab, R. D. Wells, Metal ions cause the isomerization of certain intramolecular triplexes. *J. Biol. Chem.* **267**, 1259–1264 (1992).
64. S. L. Broitman, H-DNA:DNA triplex formation within topologically closed plasmids. *Prog. Biophys. Mol. Biol.* **63**, 119–129 (1995).
65. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
66. P. Rice, I. Longden, A. Bleasby, EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
67. K. Maekawa, S. Yamada, R. Sharma, J. Chaudhuri, S. Keeney, Triple-helix potential of the mouse genome. Gene Expression Omnibus (GEO). https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE197669. Deposited 1 March 2022.