

# Topoisomerase II contributes to DNA secondary structure-mediated double-stranded breaks

Karol Szlachta<sup>1,†</sup>, Arkadi Manukyan<sup>1,†</sup>, Heather M. Raimer<sup>1,†</sup>, Sandeep Singh<sup>1</sup>, Anita Salamon<sup>1</sup>, Wenying Guo<sup>2</sup>, Kirill S. Lobachev<sup>2</sup> and Yuh-Hwa Wang<sup>1,\*</sup>

<sup>1</sup>Department of Biochemistry and Molecular Genetics, University of Virginia, Charlottesville, VA 22908-0733, USA and <sup>2</sup>School of Biological Sciences and Institute for Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received April 21, 2020; Revised May 20, 2020; Editorial Decision May 23, 2020; Accepted June 01, 2020

## ABSTRACT

**DNA double-stranded breaks (DSBs) trigger human genome instability, therefore identifying what factors contribute to DSB induction is critical for our understanding of human disease etiology. Using an unbiased, genome-wide approach, we found that genomic regions with the ability to form highly stable DNA secondary structures are enriched for endogenous DSBs in human cells. Human genomic regions predicted to form non-B-form DNA induced gross chromosomal rearrangements in yeast and displayed high indel frequency in human genomes. The extent of instability in both analyses is in concordance with the structure forming ability of these regions. We also observed an enrichment of DNA secondary structure-prone sites overlapping transcription start sites (TSSs) and CCCTC-binding factor (CTCF) binding sites, and uncovered an increase in DSBs at highly stable DNA secondary structure regions, in response to etoposide, an inhibitor of topoisomerase II (TOP2) re-ligation activity. Importantly, we found that TOP2 deficiency in both yeast and human leads to a significant reduction in DSBs at structure-prone loci, and that sites of TOP2 cleavage have a greater ability to form highly stable DNA secondary structures. This study reveals a direct role for TOP2 in generating secondary structure-mediated DNA fragility, advancing our understanding of mechanisms underlying human genome instability.**

## INTRODUCTION

DNA double-stranded breaks (DSBs) can arise during DNA metabolic processes and/or from responding to a wide range of stresses. When unrepaired or illegitimately repaired, DSBs contribute to the formation of gene rear-

rangements, deletions, and amplifications resulting in human genome instability. These modifications of the genome can introduce genomic diversity and impact evolutionary outcomes (1), however, disease-causing mutations generated by these changes involving tumor suppressor genes or oncogenes could lead to cancer development (2).

While a substantial amount of work has shown DNA repair and cell cycle checkpoint proteins to be vital for maintaining genome stability (3), alternative DNA secondary structures, which vary from the B-DNA conformation, have been demonstrated to lead to DSBs (4). DNA secondary structure-forming sequences are often found at chromosomal fragile sites (5,6). We have shown that aphidicolin-induced common fragile sites are predicted to form more stable DNA secondary structures and with greater density than non-fragile regions (7). Using DNA secondary structure calculation programs (Mfold and ViennaRNA with DNA thermodynamic parameters), we have shown that the potential for DNA stem-loop structure formation is prevalent throughout the human genome (8). Formation of these structures can occur in single-stranded DNA when the DNA duplex is unwound during processes such as replication and transcription, and can thus be influenced by nucleotide sequence and cellular activities. Once formed, stable DNA secondary structures can present a barrier for polymerase progression, resulting in incomplete replication at fragile sites and ultimately leading to DNA breakage (9). This provides a passive role for the involvement of DNA secondary structure in the initiation of DNA breaks. Several studies showed that DSBs can also occur during active transcription (10–14), and we found that DNA stem-loop structure formation is significantly enriched at transcription start sites (TSSs) (8). Canela *et al.*, Gothe *et al.* and Gittens *et al.* have recently demonstrated the involvement of topoisomerase II (TOP2) in the generation of DSBs at highly expressed genes (13–15). However, the genome-wide analysis of DSBs with respect to sites of secondary structure-forming potentials and whether there is any correlation of

\*To whom correspondence should be addressed. Tel: +1 434 243 2785; Fax: +1 434 924 5069; Email: yw4b@virginia.edu

†The author wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

TOP2-induced DSBs with the ability to form DNA secondary structure has not been explored.

DNA TOP1 and TOP2 play a critical and broad role in maintaining chromosome structural integrity during DNA processes in which strand separation generates DNA supercoiling (16,17). TOP1 and TOP2 alleviate DNA topological problems by transiently inducing a covalently-bound DNA break (single-stranded for TOP1 and double-stranded for TOP2), facilitating DNA strand passage, and then religating at the cleavage site. Recently, Hoa *et al.* (18) revealed that TOP2 frequently fails to re-ligate the endogenous, transiently-cleaved products even without the presence of inhibitors, which can therefore be processed into persistent DSBs. These persistent DSBs can occur when either a covalently bound pair of TOP2s are both processed by DNA repair machinery to result in free DNA ends, or when two single-stranded breaks occur in close proximity to one another (potentially from the action of two separate TOP2 activities occurring on opposing strands and processed to free ends). TOP2 covalently bound to DNA can be repaired by both non-homologous end-joining (NHEJ) and homologous recombination (HR) repair pathways. To undergo NHEJ repair, the covalently bound TOP2 is first degraded by the proteasome and then the remaining tyrosine-linked end can be freed by TDP2 (19–24). To undergo HR, or other pathways involving end resection, the MRN complex in cooperation with other repair proteins (such as BRCA1 and CtIP) can directly cleave off a small segment of the DNA end containing the covalently bound TOP2. This process then leaves free DNA ends that can undergo continued resection and eventual repair (25–29). The conversion of the covalently bound TOP2 DNA end to a protein free end creates the persistent break that then needs repair through these DNA repair pathways, compared to the transient nature of the TOP2 bound ends which if allowed to complete the catalytic cycle would then re-ligate the DNA ends faithfully (24).

Both TOP1 and TOP2 have very loose recognition sequences and DNA supercoiling is not an absolute requirement for TOP1 and TOP2 cleavage (22,30,31). Interestingly, several studies demonstrated a property of TOP1 and TOP2 to recognize and preferentially cleave DNA at regions capable of forming stable DNA secondary structures (32–37), similar to those predicted or formed at fragile sites. Additionally, some features found in DNA secondary structures, such as mismatched bases, are known to poison TOP2 similar to chemotherapeutic drugs (38,39). This possibility leads to a very interesting notion—whether TOP1 and TOP2 recognize non-B-DNA structural features, in addition to DNA supercoiling, and contributes to DNA secondary structure-mediated DNA fragility.

Here, we performed a non-biased, comprehensive study to evaluate the connection between DNA secondary structure formation and DNA fragility on a genome-wide scale. We further investigated the processes/proteins involved in secondary structure-mediated DNA fragility, which lead to the focus on TOP2. We found that DSBs are enriched in the regions predicted to form highly stable DNA secondary structure at a genome-wide level, and this DSB enrichment in a subset of sites is conserved across different cell types.

Stable DNA secondary structures are indeed detected in these regions in human cells, and their locations validate the computational structure predictions. Furthermore, using human DNA fragments with distinct secondary structure folding capability, we demonstrated that highly stable DNA secondary structure sequences induce fragility in yeast and display more insertions and deletions in the human genome, and the extent of instability in both analyses is in concordance with the ability to form secondary structure. In addition, TOP2 contributes to the formation of endogenous DSBs, which are enriched at the highly stable DNA secondary structure regions, and the enrichment increases upon increasing concentration of etoposide (a TOP2 inhibitor that prevents the re-ligation activity). The direct involvement of TOP2 at these structural regions was demonstrated in TOP2-deficient yeast and human cells, in which DSBs at high propensity structure regions are significantly reduced compared to TOP2 wild-type cells. The DSB regions sensitive to TOP2B knockout are energetically favorable to form secondary structure and the predicted structures are located at break summits. Overall supporting the contribution of TOP2 in DNA secondary structure-mediated breaks.

## MATERIALS AND METHODS

### Genome-wide DNA secondary structure prediction

Using DNA secondary structure calculation programs, we have provided an energetic potential for secondary structure formation across the human genome (build GRCh37/hg19) (8). Here, we applied the same analysis (ViennaRNA programs with parameters for analyzing DNA (40)), to the available sequence in the genome assembly GRCh38/hg38. With a window of 300 nt and a step of 150 nt, a total of 19 490 090 segments were analyzed, and the average  $\Delta G$  of DNA secondary structure formation was  $-30.1 \pm 11.8$  kcal/mol. Using a threshold that we identified (7), we defined sites of highly stable DNA secondary structure as having at least seven consecutive windows with a  $\Delta G$  value in the top 5% most stable structures predicted across the genome (below  $-51.3$  kcal/mol). The 23 331 sites were identified, with sizes ranging from a minimum of 1200 nt up to over 20 000 nt in length (ordered by size in Figure 1B, the leftmost panel), and these sites consist of  $\sim 1.5\%$  of the genome, indicating the stringency of the threshold.

### Cell culture and treatments

HeLa cells (ATCC) and GM13069 cells (ATCC) were grown in DMEM (Gibco) and RPMI 1640 medium (Gibco), respectively, and supplemented with 10% fetal bovine serum. Human neural progenitor cells (NPC9429) derived from an apparently healthy individual (41), were grown in DMEM/F12 (Gibco) media containing 1% N-2 (Gibco), 2% B27-A (Gibco), 1  $\mu\text{g}/\text{mL}$  laminin (Life Technology) and fibroblast growth factor (Peprotech), and maintained in matrigel (Corning)-coated dishes. Etoposide (Sigma) treatment of GM13069 cells were at 0.15, 1.5 or 15  $\mu\text{M}$  for 24 h, along with untreated cells.

### Yeast GCR assay

The yeast GCR assay has been used previously (42–44) and is based on the loss of *CAN1* and *ADE2* genes located on chromosome V. *ADE2* is placed between *CAN1* and the sequence insertion site to help differentiate between DSB-mediated arm loss coupled with repair of the broken end and mutations in *CAN1*. Mutations in *CAN1* give rise to canavanine-resistant white colonies while GCR isolates resulting from DSBs are canavanine-resistant red colonies. The spontaneous rate of GCRs in the control strains (without insert) is low:  $3 \times 10^{-9}$ /division. Each selected human sequence (AT30, chr 9: 129 987 361–129 988 350, 990 bp; AT54, chr 20: 55 532 105–55 533 500, 1396 bp; AT80, chr 11: 103 024 986–103 026 515, 1530 bp; AT85, chr 6: 61 954 246–61 955 895, 1650 bp) was PCR amplified with two sets of hybrid primers containing homology to *LYS2*. Using delitto perfetto approach (45), human motifs were inserted into chromosome V in two different orientations with respect to the direction of replication. Top2-1 allele was introduced into the AT80-containing strains via ‘pop in – pop out’ techniques using yIP5 top2-1 plasmid (46). The cells were grown on YPD at 30°C (at semi-permissive temperature for the top2-1 strains). The rates and 95% confidence intervals of the GCRs was estimated in fluctuation tests using at least 14 independent cultures.

### Indel analysis using the 10 000 human genome database

Indels were manually retrieved from the 10 000 human genome database (47). The 2 kb regions contain AT30 (chr 9: 129 986 800–129 988 800), AT54 (chr 20: 55 531 800–55 533 800), AT80 (chr 11: 103 024 750–103 026 750), AT85 (chr 6: 61 954 000–61 956 000) were queried for variations within 100 bp windows (due to database browser limitations). Next, each variation, except SNPs, were classified as an insertion or a deletion, and the length of the aberration was calculated based on the sequence. Resulting data were used in further analysis.

### Genome-wide break mapping and sequencing

Detection of DNA breaks (DSBCapture) was performed as previously described (48). Briefly, fixed nuclei of HeLa, GM13069 and NPC9429 were subjected to blunting/A-tailing reactions, and Illumina P5 adaptor ligation to capture broken DNA ends. Genomic DNA was then purified and fragmented by sonication, and subsequently ligated to Illumina P7 adaptor, and the libraries were PCR-amplified for 15 cycles. Prepared libraries were then subjected to whole-genome, 75-bp and 150-bp paired-end sequencing with the Illumina NextSeq 500 and HiSeq X Ten platform, respectively.

DNA break mapping with purified genomic DNA from GM13069 was described (49). Genomic DNA from etoposide-treated and untreated GM13069 cells was purified by gently lysing cells in 50 mM Tris–HCl (pH 8.0), 100 mM EDTA, 100 mM NaCl, 1% SDS, 1mg/mL Proteinase K for 3 h at 55 °C followed by organic extraction purification and ethanol precipitation. Precaution such as gentle pipetting with wide-opening pipette tips to avoid/minimize shearing DNA was taken to avoid introduction of DNA

breaks during purification. Purified genomic DNA was subjected to blunting/A-tailing reactions, Illumina P5 adaptor ligation to capture broken DNA ends, and subsequent reactions as described above. Genome-wide break mapping/sequencing data between experiments performed in nuclei and using purified genomic DNA from GM13069 cells are compared and discussed in the main text (see ‘Results’).

### Processing of DSB reads

Sequencing reads were aligned to the human genome (GRCh38/hg38) with bowtie2 (v.2.3.4.1) aligner running in high sensitivity mode (-very-sensitive). Restriction on the fragment length from 100 nt to 2000 nt (-X 2000 -I 100 options) was imposed. Unmapped, non-primary, supplementary and low-quality reads were filtered out with SAMtools (v. 1.7) (-F 2820). Further, PCR duplicates were marked with picard-tools (v. 1.95) MarkDuplicates, and finally, the first mate of non-duplicated pairs (-f 67 -F 1024) were filtered with SAMtools for continued analysis. For each detected break, the most 5′ nucleotide of the first mate defined the DNA break position. Sequencing and alignment statistics for the DSB mapping/sequencing libraries prepared from HeLa, GM13069 and NPC9429 cells are listed in Supplementary Table S1. Biological duplicates of each sample (N1 and N2, Supplementary Table S1) which showed very high reproducibility of genomic coverage (Pearson’s correlation  $r \geq 0.85$ ,  $P \cong 0$ ), are combined for downstream data analysis. This strong correlation confirms that the break mapping procedure does not introduce significant amounts of random DNA breaks which could convert single-stranded nicks into DSBs.

### Processing of Raji ssDNA data

The publicly available data for Raji ssDNA sequencing (SRA072844) (50) was downloaded and aligned to the GRCh38/hg38 genome using bowtie2 (v 2.3.4.1). The processed data was used to generate heatmaps based on the 23 331 highly stable DNA secondary structure regions (described in ‘heatmaps and average plots’), compute frequency among the highly stable structure, flanking, and random regions, and generate the bar graph of genomic feature distribution (described in ‘Genomic region annotation’).

### Processing of RNA-seq data

The publicly available data for GM12878 RNA-seq (SRA-SRR1153470) (51) was downloaded and aligned to the GRCh38/hg38 genome using HISAT2 (52) aligner, and the gene expression (FPKM values) were quantified using StringTie (53). In individual analyses, the expression of the genes was used to define different numbers of bins for further analysis.

### Processing of ChIP-seq data

The publicly available ChIP-seq data for HeLa CTCF (ENCSR000AKJ) (54), GM12878 CTCF (ENCSR000AKB) (54), MCF10A TOP2B (SRR5136803)



(55), and each associated input data were downloaded and aligned to the GRCh38/hg38 genome using bowtie2 (v 2.3.4.1). Binding peaks were called by the macs2 (2.1.1.20160309) as described below. In individual analyses the peak strength as defined by macs2 was used to define different numbers of bins for further analyses.

### Peak calling

The macs2 (2.1.1.20160309) software tool was used to call peaks. For ChIP-seq data macs2 was run with default settings with each dataset controlled for with the matching input data. Peak summits were then used to center the regions of interest in all other analyses.

For peak calling of GM13069 etoposide-treated samples (PRJNA497476) (49), no input data was used and a no-shift model was employed because the break data is defined by the 5' end of read 1 alone. The total of 18 791 unique break peaks were merged for the average plots, heatmaps, and box plots.

### Heatmaps and average plots

Heat maps and accompanying average plots were generated using ngs.plot.r (v 2.61). For all heatmaps using the 23 331 highly stable DNA secondary structure peaks, the reference bedfile (-E bedfile) was sorted from longest to shortest region, and this order was maintained in the heatmap (-GO none). For heatmaps for the etoposide merged peaks ( $n = 18\ 791$ ), because the region was scaled to the same size for all regions, the order for the heatmap was set to be based on the untreated sample strength of coverage in each region (-GO total) and the scale was set consistent across all plots (-SC global).

### Correlation plots

The human genome build GRCh38/hg38 was binned into 100 kb windows using BEDtools makewindows tool. The coverage in each 100 kb bin was calculated for the bam file of each replicate using BEDtools coverage ( $n = 30\ 895$ ). The read coverage in each bin is normalized to total read number (reads per million, RPM), then the absolute difference in coverage in each bin between replicates is calculated, and the top 0.05% most different (outliers) were removed. Bins where both replicates showed zero coverage were removed. Finally, data was read into python3, read normalized coverage was plotted between the two replicates, and Pearson correlation was calculated.

### Genomic region annotation

To assign genomic annotations BEDtools (v. 2.27.1) intersect was used to sequentially assign genomic features with each read only being assigned to one genomic feature. The sequential feature assignment filters out reads as they are assigned to a feature. The sequence for checking features was TSS, promoter, TTS, gene body, and those not assigned to any of these features is coded as intergenic. CTCF was separately assessed because it overlaps all of the other genomic features. The number of regions overlapping each genomic

region was then normalized to the size of the genomic region in megabases. The GRCh38/hg38 build RefSeq genes were downloaded from the UCSC browser. The definitions used for each genomic feature is as follows: promoter region ranging from TSS -1000 nt to -250 nt, TSS region ranging from TSS -250 to +250 nt, gene body region ranging from TSS +250 nt to TTS -250 nt, and TTS regions ranging from TTS -250 nt to +250 nt. The CTCF annotation used the HeLa CTCF data and defined CTCF regions as the peak summit -2000 nt to +2000 nt.

### Single-nucleotide cumulative plots around TSS and CTCF

To analyze DSBs located at the TSS regions of genes with high and low expression levels and at strong and weak CTCF-binding sites, GM12878 RNA-seq (SRA-SRR1153470) and CTCF ChIP-seq (Encode Project ENCSCR000AKB) was used, because both GM13069 and GM12878 are non-malignant lymphoblastoid cells. Based on the expression value of the genes, the TSS positions of the highest (top 10%) and lowest (bottom 10%) expressed genes were extracted ( $n = 2542$  each). The strongest (top 10%) and weakest (bottom 10%) CTCF-binding sites were determined based on macs2 score ( $n = 4019$  each) of CTCF ChIP-seq data. DSB coverage in these regions was determined using BEDtools coverage reporting the depth at each position in the reference regions (-d), and then the merge function was used to compile each region's coverage into a single line readable to python3. Using Python3 (v. 3.6.5) with matplotlib (v. 2.2.2), numpy (v. 1.15.0) and pandas (v. 0.23.3), the cumulative single-nucleotide break profiles were plotted over the relative nucleotide position to the TSS or the CTCF ChIP-seq peak summit, and in the  $\pm 2000$  bp flanking regions with read normalization (reads per million, RPM). For TSS plots the direction of transcription (on the positive or negative strand) was considered to orient the DSB profiles all in the same direction.

### Genome track images

Genome track images were made by using igvtools (v. 2.3.68) count with the options to have windows of 5 bp and precompute only 5 bp (-w 5 -z 5) for the GRCh38/hg38 build of the human genome. The resulting tdf files were loaded into the IGV browser, with the track normalization function checked in the track options to read-normalize the data, and break data tracks were set to group auto-scale. Images were then saved out from the current IGV browser view.

### Analysis of CC-seq data

The CC-seq data from Gittens *et al.* (15) was downloaded as fastq files from (GSE136943) and then aligned to the human genome (build GRCh38/hg38) following the same processing as break data (as detailed above in 'Processing of DSB reads'). The combined data from four replicates of VP16 (etoposide)-treated WT RPE-1 cells was merged into one bam file using samtools merge, and peaks were called (as detailed above in 'Peak calling') ( $n = 65\ 989$ ).

These peaks represent sites across the genome where active TOP2B enzymatically cleaves the DNA. The matched sets of VP16-treated WT and TOP2B<sup>-/-</sup> RPE-1 cells in both asynchronous and G1 arrested cells had replicates merged, respectively, and the coverage from each was calculated in the active TOP2B enzymatic cleavage peaks.

The ‘strength’ of the TOP2B enzymatic cleavage site was defined as the difference in coverage between WT and TOP2B<sup>-/-</sup> cells (strength = TOP2B<sup>-/-</sup> - WT), where the strongest sites were those with the most negative values indicating a more severe loss of activity in the knockdown line. This strength value was used to bin the peaks into 20 groups. For all sites the potential free energy profile was determined in the region  $\pm 500$  bp around the peak summit using a 300-nt window with 1-nt steps and the DNA parameters for ViennaRNA. The resulting free energy profiles were either plotted as average of medians for the entire strength bin to plot free energy compared to site strength, or the entire profile was plotted for four example strength bins.

### Statistical analysis

Statistical tests were performed using Python3 (v. 3.6.5) with scipy (v. 0.19.1) and R (v. 3.4.3). Specific tests used are detailed in figure panels and text.

## RESULTS

### Endogenous DSBs are enriched in the genomic regions predicted to form highly stable DNA secondary structures

To directly evaluate whether DNA fragility correlates with the propensity to form highly stable DNA secondary structure, we identified endogenous genome-wide DSBs at single nucleotide resolution and mapped these breaks to the genomic regions predicted to form highly stable secondary structure. Using DNA secondary structure calculation programs, we have provided an energetic potential for secondary structure formation across the human genome (build GRCh37/hg19) (8). Here, we applied the same analysis (ViennaRNA programs with parameters for analyzing DNA (40)), to the available sequence in the genome assembly GRCh38/hg38 (Figure 1A). Sliding windows of 300-nt with a 150-nt step were used across the GRCh38/hg38 assembly, and energetic potential (as  $\Delta G$ ) was assigned to each 300-nt segment. A total of 19 490 090 segments were analyzed, and the average  $\Delta G$  of DNA secondary structure formation was  $-30.1 \pm 11.8$  kcal/mol. Based on previous studies of the DNA secondary structures known to interfere with cellular processes (7,56), the highly stable DNA secondary structure site was defined as having seven or more consecutive segments with a  $\Delta G$  value within the top 5% most stable secondary structure potential ( $\Delta G$  below  $-51.3$  kcal/mol). Using this threshold, 23 331 sites that are predicted to form highly stable DNA secondary structures were identified with sizes ranging from a minimum of 1200 nt up to over 20 000 nt in length (ordered by size in Figure 1B, the leftmost panel). These sites consist of  $\sim 1.5\%$  of the genome, indicating the stringency of the threshold that we used. Among these sites, 6086 are found overlapping with a TSS  $\pm 250$  nt, and 9960 and 5765 sites exist in gene bodies and intergenic regions, respectively. When normalized to

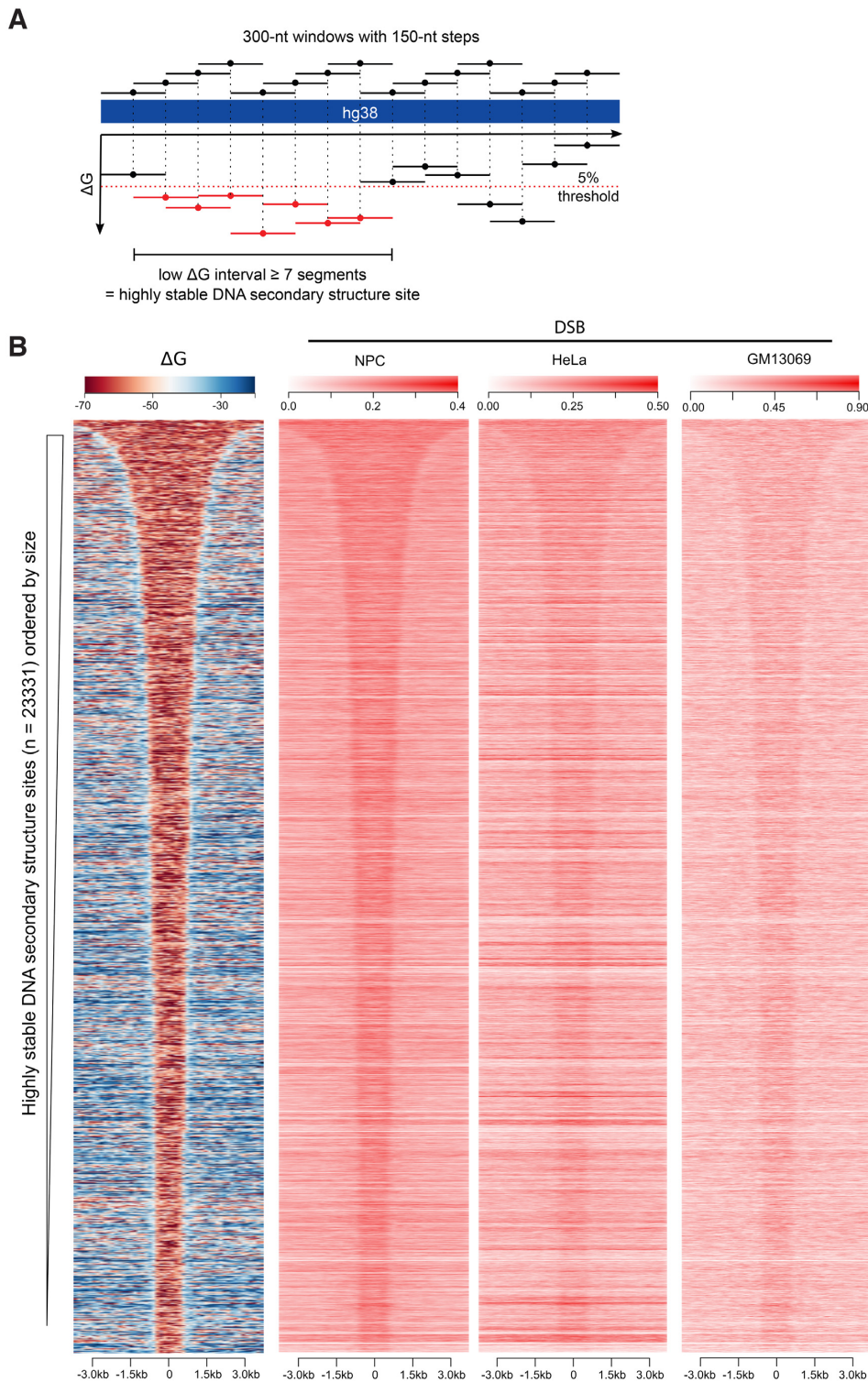
the total size of each class of genomic regions, highly stable secondary structure sites are greatly enriched at TSSs (Supplementary Figure S1, further discussion below).

To determine the propensity of these regions to break in the genome, we carried out the genome-wide single nucleotide resolution break mapping protocol, DSBcapture, adapted from Lensing *et al.* (48) (Supplementary Figure S2) in three different cell types, human neural progenitor cells (NPC) derived from an apparently healthy individual, HeLa, and non-malignant human lymphoblastoid cells, GM13069 (Supplementary Table S1). Two biological replicates for each cell type were performed, displayed high correlation of genomic coverage (Pearson’s correlation  $r \geq 0.85$ ,  $P \cong 0$ , Supplementary Figure S3A), and pooled for analysis. DSB coverage over the 23 331 sites of highly stable DNA secondary structure showed that DNA DSBs are enriched at those sites in all three cell types (Figure 1B). The DSB enrichment at the stable DNA secondary sites was observed in each replicate as well (Supplementary Figure S3B). Additionally, analyzing previously published DSB data (48) of DSBs mapped in normal human epidermal keratinocytes (NHEK) demonstrated a similar enrichment of breaks in the highly stable secondary structure sites compared to flanking regions (Supplementary Figure S4).

To evaluate whether the enrichment of DSBs at the highly stable structure sites (low  $\Delta G$  intervals) was meaningful, the DSB coverage at these sites was compared to those at sites generated by random shuffling (random) and to 1 kb regions flanking the highly stable structure sites (flanking). In all three cell types examined, DSB coverage is significantly higher within the highly stable secondary structure sites than within the two control groups ( $P \cong 0$ , one sided Mann-Whitney U test) (Supplementary Figure S5). These results suggest that genomic regions with a high potential to form stable DNA secondary structure are more prone to DNA breakage compared to both random and flanking sequences. Furthermore, the similarity in DSB enrichment across the tested cell types at these sites of highly stable secondary structure argues for a contribution of DNA secondary structure in the promotion of DNA fragility irrespective of cell specificity.

### DNA secondary structures experimentally identified in human cells validate computational structure prediction

Genome-wide non-B DNA secondary structures have been recently, experimentally determined in Raji cells, a human lymphoblastoid cell line, using the combination of permanganate footprinting, high-throughput sequencing, and the sequence motifs of non-B DNA (50). This method utilized the ability of permanganate to modify exposed DNA bases and prevents the re-annealing of these modified bases. The inability of the exposed and modified bases to re-anneal then made the sites vulnerable to mung-bean nuclease digestion to generate DNA breaks which allowed for the capture and high-throughput sequencing of these sites (57). The comparison of these sequenced regions to known non-B DNA secondary structure motifs determined that these single-stranded regions were enriched for motifs of non-B DNA secondary structures (50). Therefore, to examine whether non-B DNA structures detected in cells are present



**Figure 1.** Highly stable DNA secondary structures (low folding  $\Delta G$ ) correlates with elevated DSBs. **(A)** A schematic depicting how highly stable DNA secondary structure sites were selected. DNA secondary structure folding  $\Delta G$  of the human genome (GRCh38/hg38) was calculated in 300-nt windows with a 150-nt step, represented by the 300-nt bars across the top. Regions with at least seven consecutive segments having folding  $\Delta G$ s in the top 5% most stable structures predicted across the genome (5% threshold indicated by the red dashed line) were defined as highly stable DNA secondary structures (low  $\Delta G$  intervals, shown in red). Dots indicate the center of each 300-nt segment to which the calculated  $\Delta G$  is attributed. **(B)** Heatmap representations of folding  $\Delta G$ s (left panel) and DSB coverage in three cell lines, human neural progenitor cells (NPC), HeLa and non-malignant human lymphoblastoid cells, GM13069, are shown in the 6 kb regions centered on the highly stable DNA secondary structure intervals (ordered by size) as defined in panel A. Heatmap of highly stable secondary structure interval folding  $\Delta G$ s (left panel) represents the potential folding energy of each 300-nt segment (as shown in panel A) within the interval (ranging from 1200 to 20 000 nt in size) and the flanking region for the 23 331 identified regions. DSB coverage heatmaps presents the read coverage across the stable secondary structure intervals and associated flanking regions.



in the regions predicted to form highly stable secondary structures, we compared our computationally identified highly stable secondary structure regions with the genome-wide measurement of non-B DNA secondary structures in Raji cells. We found that the secondary structures detected in Raji cells are preferentially enriched in these regions (Figure 2A), and the predicted regions of highly stable secondary structure contain significantly more detectable non-B form structures than the randomly shuffled regions and the adjacent regions ( $P \cong 0$ , one-sided Mann–Whitney  $U$  test) (Supplementary Figure S6). Altogether, this demonstrates that stable DNA secondary structures are present in human cells, and their locations validate our computational structure predictions. Combined with the observation in Figure 1B, we found an overall correlation among the secondary structure footprints, the predicted low  $\Delta G$  intervals, and the DSB profile, with an example of a genomic region shown in Supplementary Figure S7.

To further validate the computationally-identified highly stable secondary structure regions for structure-forming capacity in cells, the presence of G-quadruplex structures in human cells was assessed. The existence of G-quadruplexes (G4), a subtype of non-B DNA secondary structures, has already been experimentally demonstrated in NHEK cells, using an antibody specific to G4 structures followed by Illumina DNA sequencing (58). The ViennaRNA program (with thermodynamic parameters for DNA analysis) that we used for secondary DNA structure prediction can calculate the potential formation of G4 in addition to stem-loop structures. Therefore, we calculated the coverage of predicted highly stable secondary structure at the experimentally detected G4 regions, and identified 4229 intersecting regions genome-wide that are prone to form G4 structure. We then analyzed the relationship of the computed secondary structures including G4 ( $\Delta G$ ), and both G4 ChIP-seq reads and DNA breaks detected in NHEK cells (Figure 2B). We found that the distribution of G4 structures corresponds with the computed structure, indicating the presence of G4 structures at the predicted sites, thereby adding another validation of the computational secondary structure predictions. More importantly, the DSB coverage profile shows the enrichment of DSB at the 4229 predicted secondary structure sites containing the experimentally detected G4 structure.

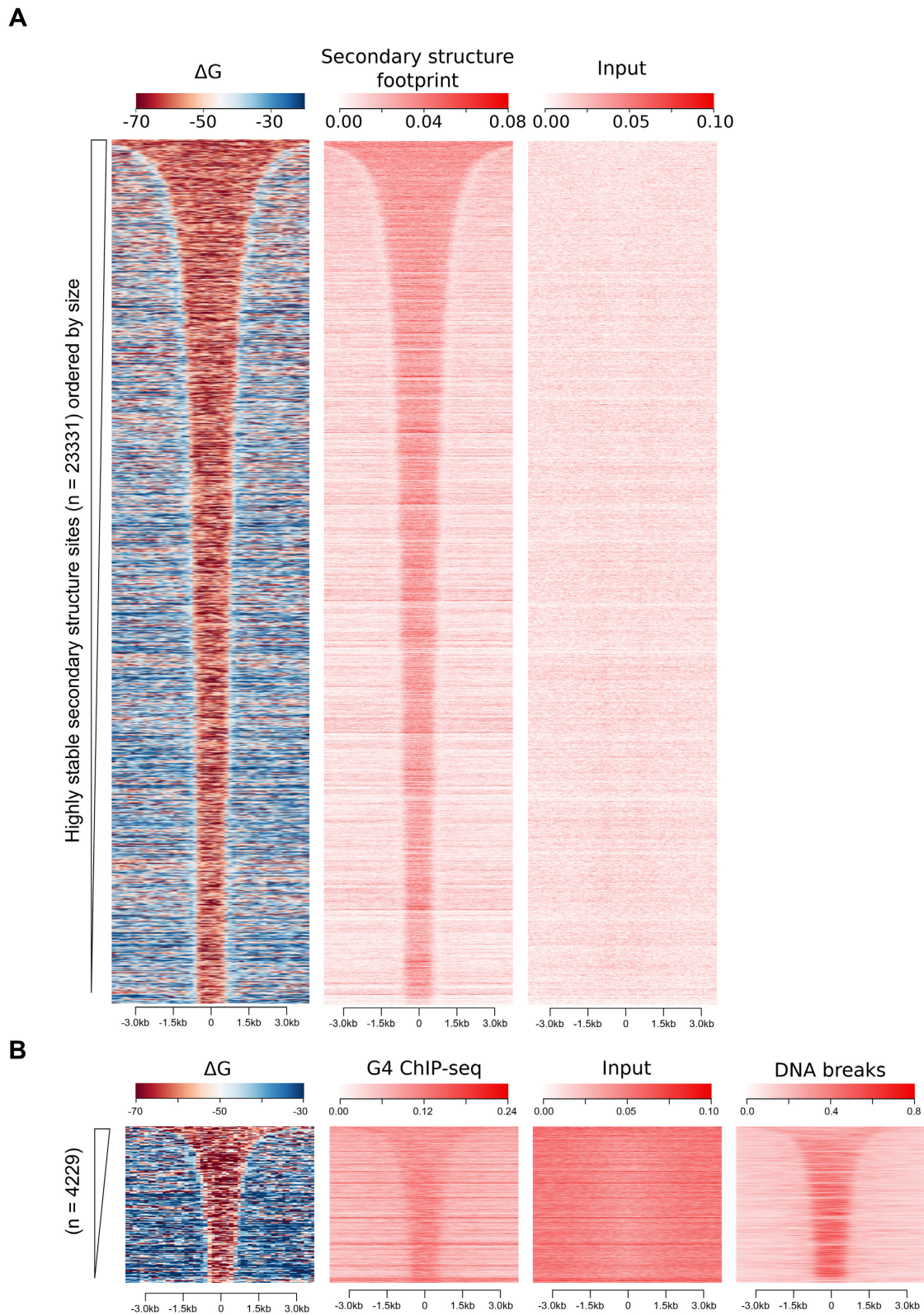
These two experimental structural analyses provide direct evidence for the existence of non-B form DNA secondary structures in human cells at computationally predicted loci, and more importantly, support the presence of DNA secondary structures at DSB sites, strongly suggesting the involvement of non-B form DNA secondary structure in DSB generation.

### Regions of highly stable secondary structure induce fragility in yeast and display instability in the human genome

To determine the fragility and instability of non-B DNA secondary structure-forming sequences, we selected sequences to test more directly. Four AT-rich (~65% AT-content) human DNA fragments with distinct potential secondary structure folding  $\Delta G$ s (Supplementary Figure S8) were selected: AT30 with an average  $\Delta G$  of  $-29.97$

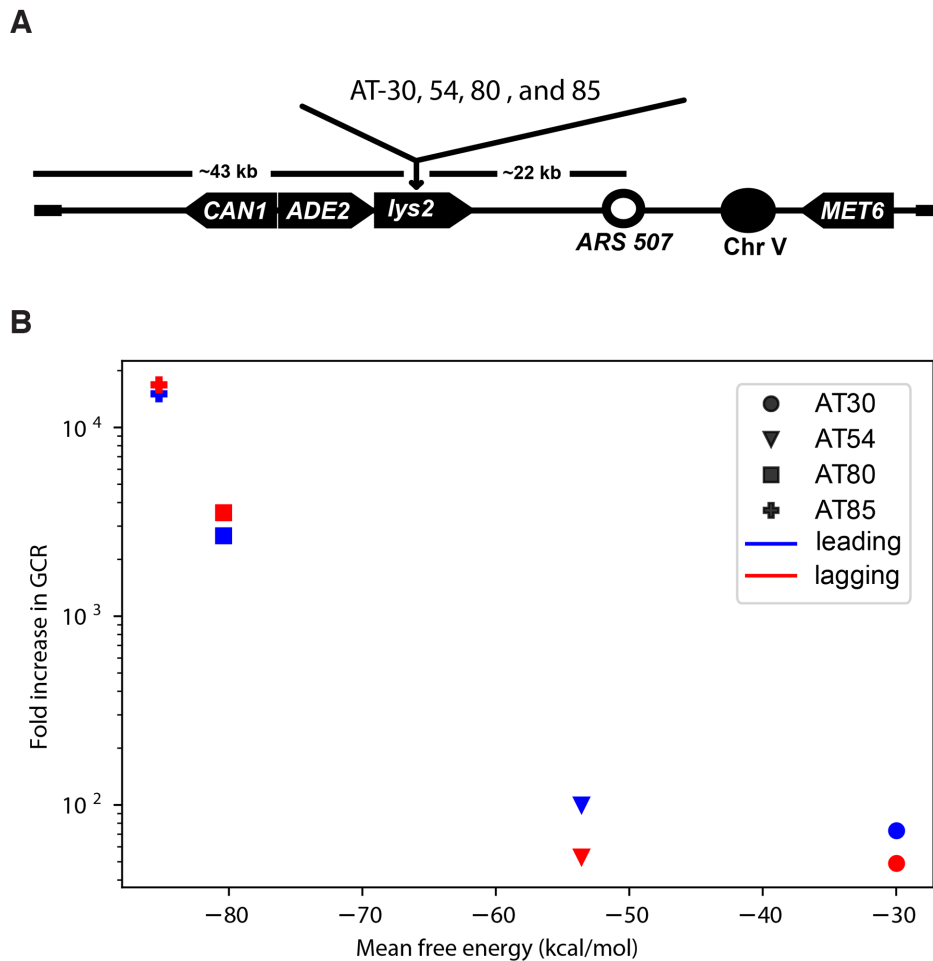
kcal/mole (similar to the average  $\Delta G$  of the human genome,  $-30.1 \pm 11.8$  kcal/mol (standard deviation, SD)), AT54 with  $-53.84$  kcal/mol (similar to the average plus 2 SDs), AT80 with  $-80.29$  kcal/mol (as the average plus 4 SDs), and AT85 with  $-85.32$  kcal/mol (Supplementary Table S2), with the length ranging from 990 to 1650 bp. These four AT-rich sequences were named in a manner that the numbers in their name represent their negative  $\Delta G$  values. The AT-rich characteristic of the fragments was chosen to emphasize the ability to fold DNA secondary structure, rather than high GC content which can potentially form multiple kinds of non-B DNA structures. To experimentally validate the potential to form secondary structure, AT30 and AT85 fragments were generated and subjected to re-duplexing assays to examine the propensity to form stem-loop structures. This *in vitro* re-duplexing assay, which allows re-annealing of the single-stranded DNA at low concentrations following denaturation, has been used to analyze the formation of DNA secondary structures generated by a variety of sequences (7,59–61). Upon separation on native polyacrylamide gels, the re-duplexed AT85 displays more slow-migrating products compared to AT30, indicating AT85 has a greater tendency than AT30 to form stable secondary structures under single-stranded conditions (Supplemental Figure S9A). The major parameter contributing to the stability of secondary structure folded from a single-stranded DNA is the extent of base-paired regions within each structure. Next, the re-duplexed AT30 and AT85 were digested with mung bean nuclease (degrading single-stranded regions) to examine the distribution of base paired regions. The structural probing experiments demonstrated that indeed AT85 contains more stems and longer stem lengths than AT30 (Supplemental Figure S9B), again validating the secondary structure prediction that AT85 is more favorable to form stable DNA secondary structure under single-stranded state. In addition, examination of all four AT-rich sequence regions for secondary structure footprinting reads in Raji cells suggests that DNA secondary structure can form at these sequences in cells.

Because of the similarity in DSB enrichment across cell types (Figure 1B and Supplementary Figures S3 and S4), we next examine the structure-driven fragility of these four AT-rich sequences in yeast to further exclude any cell type specific effects. Using the gross chromosomal rearrangement (GCR) assay in yeast (42–44), these four human DNA fragments were integrated into a *LYS2* cassette on chromosome V of the yeast genome (Figure 3A). In this assay, canavanine-resistant red colonies ( $\text{Can}^R\text{Ade}^-$ ) occur due to DSBs in the *LYS2* region and repair of the broken chromosome. Therefore, the GCR rates reflect the break frequency of the inserted DNA. All four human DNA sequences showed an increase in the GCR rates compared to the background (no insert). This induction of GCRs is proportional to increasing secondary structure-forming ability ( $\Delta G$  decreasing) (Figure 3B), with the AT85 sequence showing the highest GCR rate (>15 000-fold) (Supplementary Table S2). This demonstrates that DNA fragility, measured by the GCR rate in yeast, reflects the ability of these human sequences to form stable DNA secondary structure. Each human sequence was inserted into the chromosome V arm in two orientations relative to the replication ori-



**Figure 2.** Stable DNA secondary structures detected in human cells are preferentially enriched in the region with computed highly stable DNA secondary structures. (A) Heatmap profiles of DNA secondary structure footprint coverage in Raji cells (middle panel) are shown over the regions of highly stable DNA secondary structures (low  $\Delta G$  intervals, as defined in Figure 1A) (left panel). The input data from the secondary structure footprint study in Raji cells is also included (right panel). (B) Heatmap representation of G4 ChIP-seq (58) and DSB break (48) coverage in NHEK cells over the low  $\Delta G$  intervals that contain G4 ChIP-seq peaks.



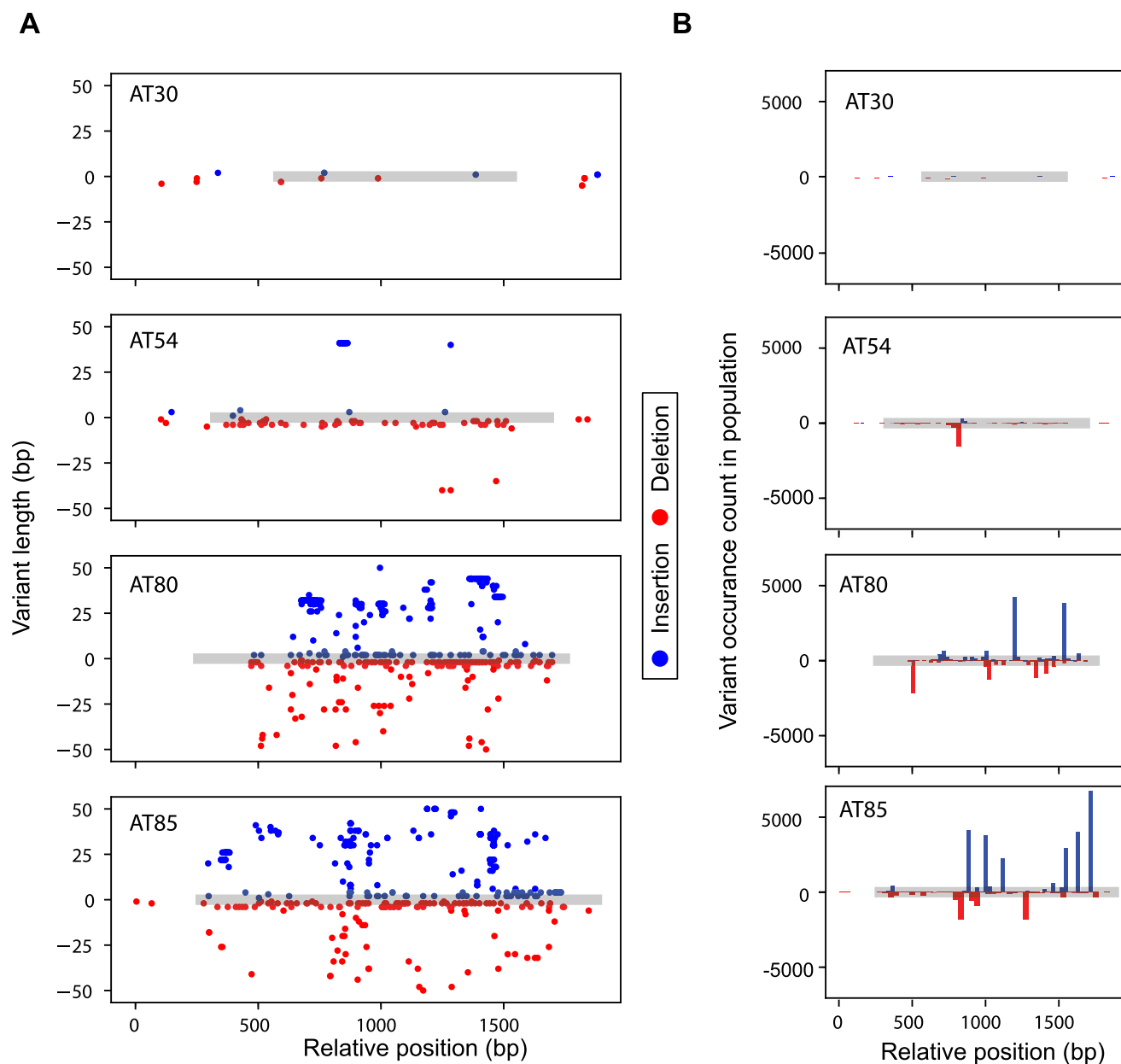


**Figure 3.** Secondary structure-rich human sequences are prone to DSBs in yeast. (A) Schematic representation of the GCR assay, based on the loss of *CAN1* and *ADE2* genes located on chromosome V. Four human DNA sequences (AT30, AT54, AT80 and AT85) with distinct secondary structure folding  $\Delta G$ s (Supplementary Table S2 and Supplementary Figure S8) were inserted into the *LYS2* gene with two orientations relative to the *ARS 507* replication origin. (B) DSB detection by GCR assay shows that DSBs proportionally increase as the ability to form secondary structure increases ( $\Delta G$  decreases) in these four AT-rich human sequences. Both GCR rates and fold increases over the background (no insert) of all sequences are listed in Supplementary Table S2.

gin *ARS 507*, and this design places the positive strand of the human sequence in either leading or lagging strand during replication from *ARS507*. Each sequence regardless of the strand location during DNA replication shows similar GCR rates, possibly because the positive and negative strands of each sequence have very similar average  $\Delta G$ s (the ability to fold secondary structure). The *LYS2* gene, in which the sequences were inserted, has an active promoter, and therefore GCR events could occur from DNA secondary structure formation during transcription of the *LYS2* gene.

We next examined whether the instability observed in the yeast genome among these four sequences correlates with genetic variations in the flanking regions for these motifs in the human population. Using the 10 000 human genome database (47), we identified sequence variants within the 2 kb regions containing AT30, AT54, AT80, and AT85 sequences for the presence of insertions and deletions (indels) and excluding point mutations. The 2 kb region containing AT80 and AT85 has a greater variety of indels (466 and 529,

respectively) than those found in the 2 kb region of AT30 and AT54 (20 and 96, respectively). Additionally, the size of indels are larger in AT80 and AT85 (both have a median indel size of 41 bp) when compared to those in AT30 and AT54 (median indel size 1 and 4 bp, respectively) (Figure 4A). To account for the abundance of variants, the number of occurrences of each indel variant, by genomic coordinate, is plotted in 50-bp bins, and the total number of indel occurrences in the 10 000 genomes cohort is higher in AT80 and AT85 than AT30 and AT54 (Figure 4B). AT30 which has an average folding  $\Delta G$  at a level of the genome average has few variants. AT54 has more indels than AT30 in types, sizes, and total numbers, but they are not of the same extent as in AT80 and AT85. In AT80 and AT85, variants are clustered within regions of very stable DNA secondary structure. These results indicate that sequences that are break-prone in yeast and have potential to form stable DNA secondary structures (AT80 and AT85) display genetic polymorphism with more types and numbers of indels and larger sizes of these indels, compared to less-structured



**Figure 4.** Analysis of indels retrieved from the 10 000 human genomes database (47) shows that the break-prone regions with stable secondary structure (AT80 and AT85) cluster with more and larger indels than less structured low-break regions (AT30 and AT54). (A) Location and length of unique insertions and deletions observed in healthy humans are shown within the 2 kb regions containing AT30, AT54, AT80 and AT85 sequences, from top to bottom respectively. (B) Occurrence of each insertion and deletion from the 10 000 genomes cohort shown in panel A are presented in 50-bp bins across the 2 kb regions. Indel frequencies ranged from 0.1% to 47% of the 10 000 genomes cohort. The AT30, AT54, AT80 and AT85 sequences are indicated within each 2 kb region (grey horizontal bars).

regions of lower breakage with a similar AT-content, suggesting a possible impact of structure-driven fragility in the human genome.

#### TOP2 contributes to DSBs at DNA secondary structure-rich regions

We next investigated which processes/proteins are involved in the generation of DNA secondary structure-mediated DNA breaks. DNA secondary structure formation is prevalent throughout the human genome, and highly stable struc-

ture sites are enriched at transcription start sites (TSSs) and CTCF-binding sites based on the computational prediction (Supplementary Figure S1) (8) and experimental probing data in Raji cells (Supplementary Figure S10) (50). TSS and CTCF-binding sites are two major TOP2 activity sites, and are known to accumulate DNA supercoiling, which invites TOP2 activity. To test whether TOP2 contributes to the generation of DSBs at these sites, GM13069 cells were treated with increasing concentrations of etoposide, an inhibitor of the ligation activity of TOP2, and DSBs were measured and compared to DSBs of untreated cells (49). To lower the

necessary starting materials, here the mapping/sequencing experiments were performed using purified genomic DNA with precautionary steps to minimize the introduction of DNA breaks (49). The reproducibility of the mapped break data between using purified genomic DNA and DSBCapture in nuclei in untreated GM13069 samples was highly correlated (Pearson's correlation  $r = 0.98$ ,  $P \cong 0$ ; Supplementary Figure S11).

We then analyzed the differences in DSB frequency between treatments across all TSSs and CTCF binding sites, each in ten bins based on expression level and CTCF binding strength, respectively (Supplementary Figure S12). A clear, strong relationship between DSB frequency and both expression level and binding strength is observed, and the increasing etoposide-induced breaks are most apparent at stronger CTCF binding sites and higher expressing TSSs. The distribution of DSB around these sites revealed that TSSs of highly expressed genes and peaks of strong CTCF-binding sites display a DSB enrichment over the background of untreated cells, and the DNA break increase corresponds to the increased concentrations of etoposide (Figure 5A and B, top panels). In contrast, neither the DSB enrichment nor the increasing etoposide-induced breaks were observed at the TSS regions of low expressed genes or among weak CTCF-binding sites (Figure 5A and B, bottom panels).

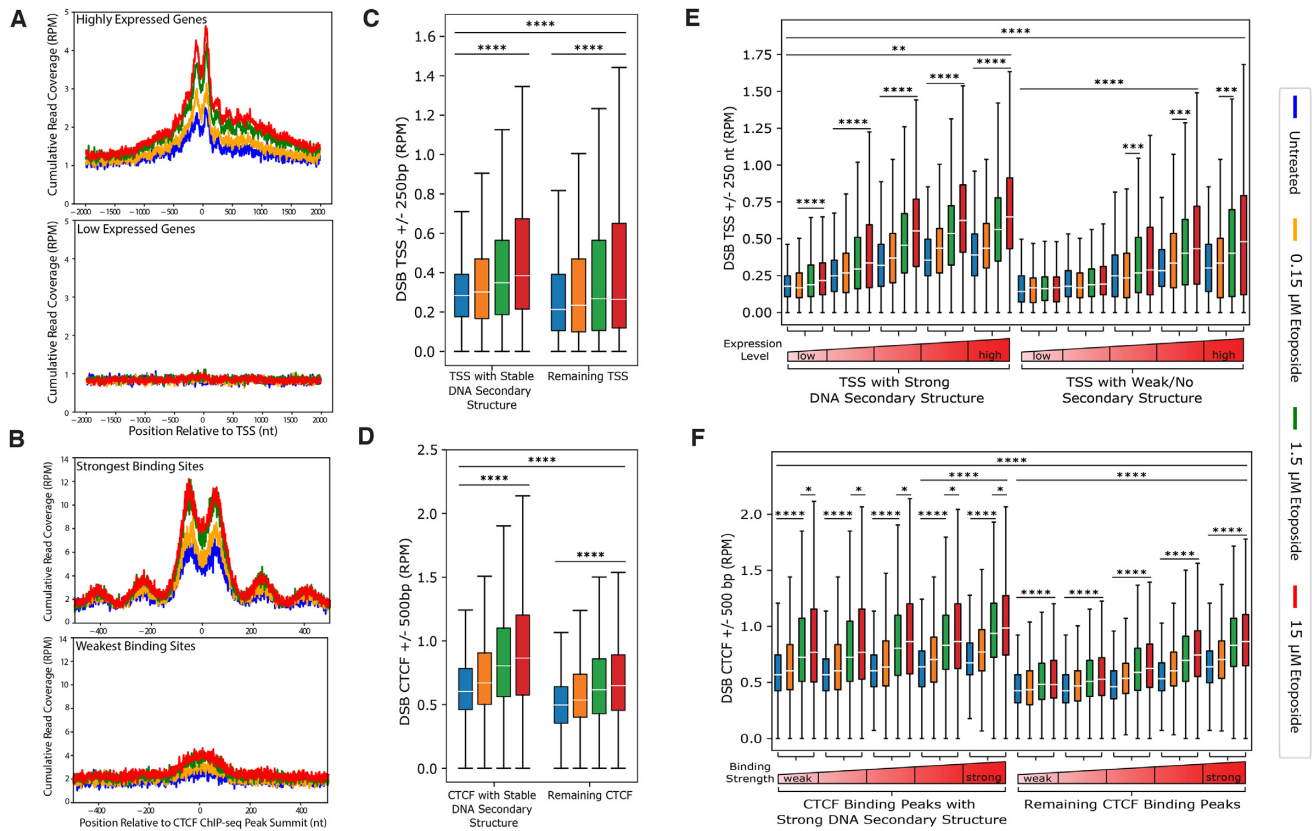
To directly assess the relationship between DNA secondary structure and DSBs at these regions, we divided all TSSs and CTCF binding sites into those that were overlapping our 23 331 highly stable DNA secondary structure regions, and those in the remaining regions. We found that the TSS and CTCF regions overlapping the highly stable regions were significantly more enriched with DSBs and maintained the strong increase in breaks upon etoposide treatment compared to those not found in the strong DNA structure regions (Figure 5C and D,  $p \cong 0$  Kruskal–Wallis with Benjamini–Hochberg corrected Dunn test). Furthermore, when TSSs and CTCF binding sites in the highly stable secondary structure and the remaining region groups were binned by either expression level or CTCF binding strength, respectively, we found that the strongest response of etoposide-induced breaks in accordance with expression and binding strength is within the strong secondary structure group (Figure 5E and F). The remaining or no secondary structure groups in TSS and CTCF still do have a response to both etoposide and expression or binding strength with respect to DSBs, but no to the same extent as those in strong secondary structure groups. These results suggest that genomic regions enriched with DNA secondary structures (such as TSS and CTCF) could generate DSBs by the action of TOP2, which can be stimulated by both gene expression levels and CTCF binding strength. One such example is the intron10/exon 11 junction of the *KMT2A* gene known to form a multiple stem-loop structure (62) and to contain rearrangement breakpoints in therapy-related leukemia patients with exposure to TOP2 inhibitors (63,64). We found that this region displays preferential DSB in untreated cells and increased DSB frequency with increasing etoposide treatment (Supplementary Figure S13), supporting the contribution of TOP2 in DNA breakage at a DNA secondary structure-rich region.

In Figure 5 A and B (top panels), we noticed endogenous DSBs in untreated cells also showed an enrichment at both TSSs and CTCF-binding sites compared to the adjacent gene regions and displayed a similar DSB distribution pattern as for etoposide-induced samples. Therefore, we next examined the endogenous DSBs and its relationship to TOP2. We found that DSBs tend to occur in clusters, with the number of identified break cluster peaks increasing upon etoposide treatment (214, 618, 9,097 and 14,961 break peaks, respectively, for untreated, 0.15, 1.5 and 15  $\mu$ M of etoposide) (Supplementary Figure S14A). Interestingly, there is a substantial overlap among the peaks identified across treatments, specifically higher etoposide concentration peaks overlap with peaks from low etoposide and untreated samples. Within the total 18 791 unique break peaks merged from all treatments, DSB coverage in these peaks increases with increasing inhibition of TOP2 re-ligation activity (Supplementary Figure S14B and C). Additionally, heatmaps of DSB coverage over these peaks (Supplementary Figure S14C) also reveal that in untreated cells, a snapshot of endogenous breaks was captured, and etoposide traps these breaks as the concentration increases.

To explore the possibility that TOP2 contributes to DNA secondary structure-mediated DSBs, the ability of these break peaks to form DNA secondary structures was analyzed, and indeed, the TOP2 break peak regions have significantly greater potential (lower  $\Delta G$ ) to form stable DNA secondary structure than the rest of the genome ( $P \cong 0$ , one sided Mann–Whitney  $U$  test) (Supplementary Figure S14D). We then asked whether the 23 331 highly stable DNA secondary structure sites (Figure 1B, the leftmost panel) are enriched with TOP2-mediated DSBs. DSB coverage across these stable secondary structure regions showed that etoposide treatment increased DSBs as compared to untreated, and the intensity of DSBs significantly increases with increasing etoposide concentrations (Figure 6), indicating that TOP2-mediated DSBs preferentially present at highly stable DNA secondary structure regions.

To assess the direct role of TOP2 in genomic fragility, we first revisited the yeast GCR assay with one of the previously tested AT-rich sequences, AT80, and introduced the temperature sensitive *top2-1* allele, which has significantly reduced enzyme activity (46) (Supplementary Table S2). The *top2-1* strain with the AT80 fragment in both the leading and lagging position has a significantly lower GCR rate compared to the wild-type strain (Figure 7A,  $P < 0.01$ , unpaired  $t$ -test), further indicating that TOP2 contributes to the genomic fragility of these non-B DNA secondary structure sequences, independent of cell type. Next, we examined the effect of TOP2B knockout in the generation of DNA breaks in human cells using the recent CC-seq data in RPE-1 cells by Gittens *et al.* (15), that maps TOP2 cleavage complex-associated DSBs at single-nucleotide resolution. We first identified DSB peaks from wild-type cells treated with etoposide, to indicate genomic regions targeted by TOP2 activity ( $n = 65\ 989$ ). Then we evaluated the etoposide-induced DSBs of paired wild-type and TOP2B knockout (TOP2B<sup>-/-</sup>) cells in both asynchronous and G1 arrested states at the identified peak regions. The DSB profile reveals that in both asynchronous and G1 states there is a significantly lower break frequency in the TOP2B<sup>-/-</sup> cells



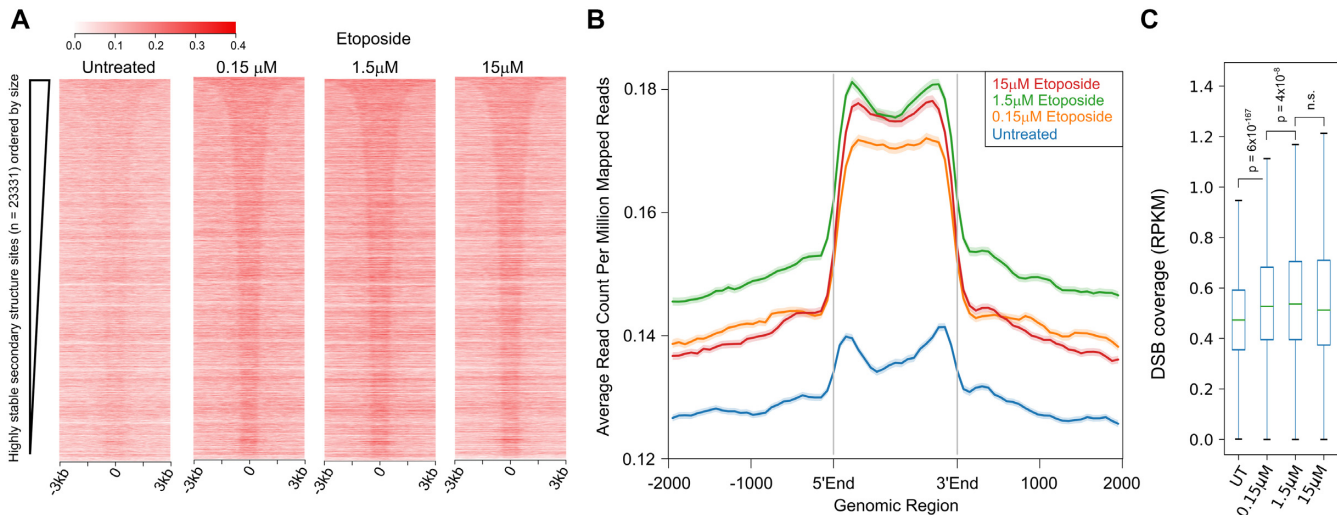


**Figure 5.** Etoposide induces DSBs at TSSs and at CTCF-binding sites, and those sites with stable DNA secondary structures are more prominent with etoposide-induced DSBs than sites of weak/no secondary structures. Cumulative, read normalized, single nucleotide resolution profiles of DSBs over TSS (A, top panel, the top 10% highly expressed genes,  $n = 2542$ ; bottom panel, the 10% least expressed genes,  $n = 2542$ ) and CTCF-binding sites (B, top panel, the top 10% strongest binding sites,  $n = 4019$ ; bottom panel, the 10% weakest binding sites,  $n = 4019$ ), present an enrichment over background in untreated cells, and the enrichment increases with increased concentrations of etoposide. (C) DSBs at TSS  $\pm 250$  bp (RPM, reads per million) are compared between TSSs intersecting with the 23 331 highly stable DNA secondary structure regions ( $n = 7350$ , 'TSS with Stable DNA Secondary Structure') and those outside these regions ( $n = 18\,067$ , 'Remaining TSS'). (D) DSBs at CTCF  $\pm 500$  bp (RPM) are compared between CTCF sites overlapping the 23 331 highly stable DNA secondary structure regions ( $n = 4127$ , 'CTCF with Stable DNA Secondary Structure') and those outside these regions ( $n = 36\,062$ , 'Remaining CTCF'). (E) The TSS regions belonging to the highly stable and the weak DNA secondary structure groups respectively, were further binned into five groups based on the gene expression level of all genes. (F) The CTCF regions belonging to the highly stable and the weak DNA secondary structure groups respectively, were further binned into five groups based on their CTCF binding strength determined by the strength of all ChIP-seq peaks. Boxes denote 25th and 75th percentiles, middle bar shows median, whiskers span from 5% to 95%, and outliers are not shown. \*  $P < 0.025$ ; \*\*  $P < 0.02$ ; \*\*\*  $P < 0.01$ ; \*\*\*\*  $P \geq 0$ , Kruskal–Wallis followed by Dunn Test post-hoc with Benjamini–Hochberg correction.

compared to the wild-type cells at these regions (Figure 7B,  $P \geq 0$ , two-sided Wilcoxon signed-rank test), demonstrating that TOP2 is involved in generating these breaks. Interestingly, the DSB reduction by knocking out TOP2B is more prominent in G1 cells versus in asynchronous cells, indicating TOP2B is a major contributor to generate cleavage complex-associated DSBs (TOP2B, the main TOP2 in G1), and in asynchronous cells TOP2A can compensate the action of TOP2B when it is knocked out. In addition, we also examined TOP2 cleavage complex-associated DSBs in the untreated G1-arrested wild-type and TOP2B<sup>-/-</sup> cells, and similarly saw a significant decrease with the TOP2B knock-out (Supplementary Figure S15,  $P \geq 0$ , two-sided Wilcoxon signed-rank test).

To understand the structure forming propensity of these TOP2 activity peaks, we separated the 65 989 peaks into twenty groups ( $n = 3283$  per group) based on the strength of TOP2 activity at the peak defined as the difference in breaks between the wild-type and knockout cells. Then

DNA secondary structure propensity was calculated and revealed that stronger TOP2 activity correlates with more stable DNA secondary structure formation (Figure 7C, Pearson's correlation  $r = 0.908$ ,  $P = 3.19 \times 10^{-8}$ ). Furthermore, we evaluated the folding energy profiles around the TOP2 activity peaks for four of the twenty groups, representing the strongest, weakest, and two intermediate strength bins. Comparing the median free energy of folding between these four sets of peaks again indicates that the most sensitive regions to TOP2 activity have a more favorable secondary structure-forming potential and that the most stable DNA secondary structures occur at cleavage complex-associated break summits, as indicated by the free energy minima centered at the peak summits (Figure 7D). Interestingly, all regions assessed have median free energies lower than the genome-wide median. Finally, we compared the CC-seq captured, TOP2-activity breaks to TOP2B ChIP-seq peaks (55) to evaluate the overlap of TOP2 binding and cleavage activity. We observed that there is a sharp, defined peak of



**Figure 6.** Etoposide-induced DSBs are enriched at the highly stable DNA secondary structure regions (low  $\Delta G$  intervals). (A) Heatmaps demonstrate DSB coverage at the highly stable secondary structure regions ordered by size as the leftmost panels in Figures 1B and 2A. (B) Mean coverage of DSBs at the stable structure regions  $\pm 2$  kb of flanking region as presented in the heatmaps in panel A. Highly stable secondary structure regions are scaled to the same size and are represented as 5' to 3' end of each individual region. (C) Box plot representation of DSB density at the highly stable DNA secondary structure regions (low  $\Delta G$  intervals) (RPKM, reads per kilobase million). DSB intensity is significantly increased with increasing etoposide concentration (one sided Mann-Whitney  $U$  test).

TOP2 CC-seq signal at the summits of TOP2B ChIP-seq peaks (Supplementary Figure S16). Furthermore, in concordance with the strong peak of TOP2 cleavage activity, these TOP2B peak summits also have a greater likelihood to form stable DNA secondary structures, indicated by the strong minima for folding energy at these peaks (Supplementary Figures S16). Overall, these results demonstrate that the regions that are particularly prone to TOP2 binding and cleavage are those that are also most likely to form highly stable DNA secondary structures, suggesting a likely role for TOP2 cleavage activity at these secondary structure forming regions.

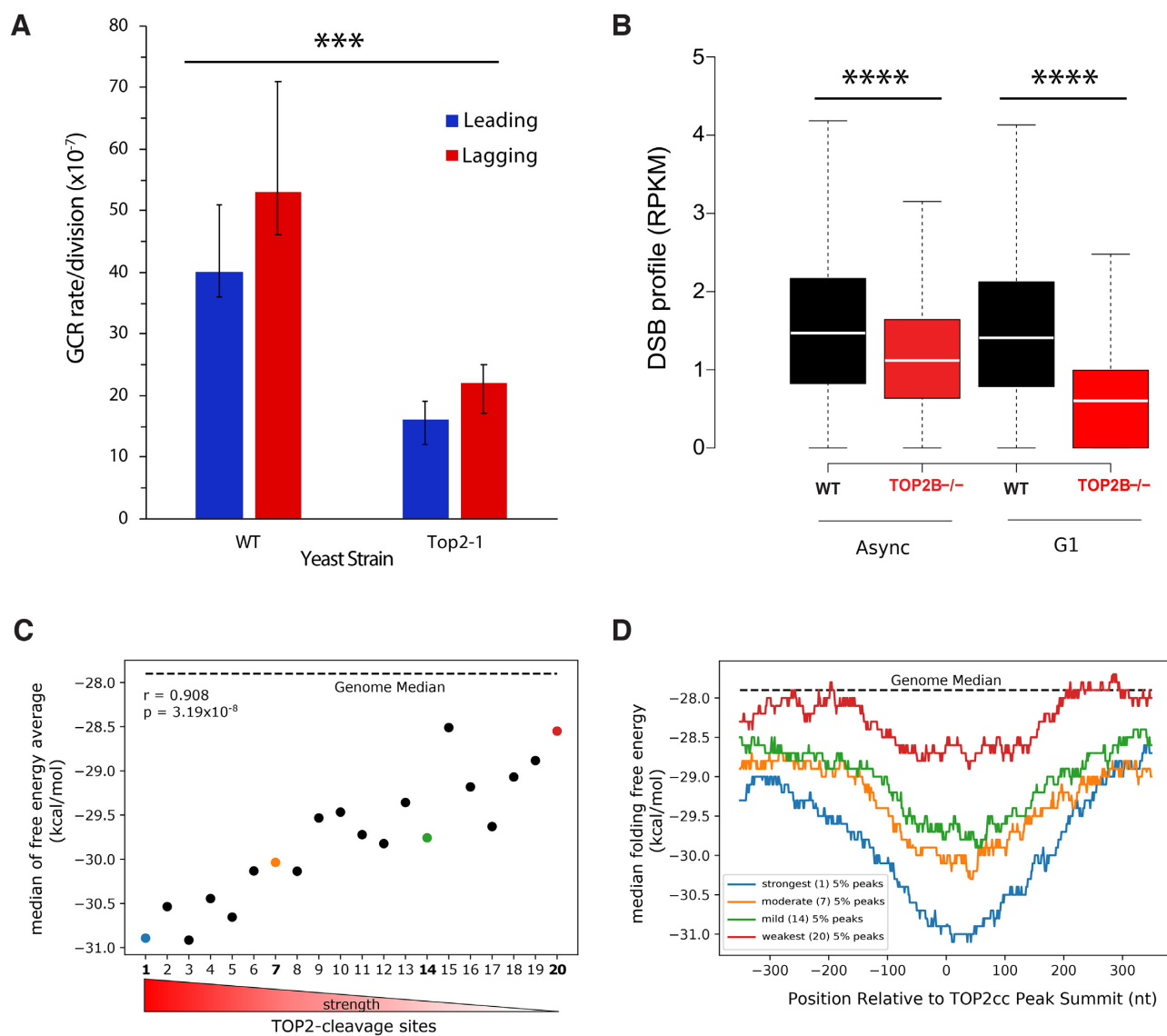
## DISCUSSION

Our genome-wide, non-biased study demonstrated endogenous DNA breaks enriched at regions forming highly stable DNA secondary structure, and a set of break-prone structure-rich regions are conserved across different cell types, supporting a structure-driven mechanism for DNA fragility. The yeast GCR assay and the indel analysis of a human population both showed that the ability of a region to form alternative DNA secondary structure is positively related with an increase in fragility/instability of the region. The observation of enrichment of alternative DNA secondary structure at TSS regions and CTCF-binding sites reveals the contribution of TOP2 to the endogenous breaks at these genomic feature regions. Etoposide-induced break sites have a higher potential to form alternative DNA secondary structure than the rest of the genome and highly stable secondary structure sites are enriched with increasing etoposide-induced breaks. Finally, top2-1 yeast GCR assays and analysis of the TOP2 cleavage complex-associated DSBs demonstrates that TOP2 has a direct role in the fragility at regions of highly stable DNA secondary structures.

The DNA secondary structure prediction program that we employed can compute the formation of hairpin, G4, and multiple stem-loop structures. In addition to identifying genomic regions with the potential to form DNA secondary structures from single-stranded DNA, the strength of this analysis is to estimate the relative propensity for these regions to actually form such structures. This allows us to discover and rank regions based on the structure-forming potential which might be overlooked when relying only on the primary sequences of the region. Studies of multiple stem-loop structures have been limited to DNA sequences at specific known loci such as fragile sites (5,6). The abundance of these break-prone structure-rich regions observed in our studies suggests DNA secondary structure-mediated DSBs as a common mechanism for DNA fragility.

Alternative DNA secondary structures have been mapped to breakpoint regions of disease-causing chromosomal rearrangement in human. The palindromic AT-rich repeats which form multiple stem-loop structures are present at translocations  $t(11;22)$  in parents of Emmanuel syndrome children (65) and  $t(3;8)$  translocation found in renal cell carcinoma (66). Polypurine or polypyrimidine sequences with mirror repeat symmetry can form intramolecular triplex known as H-DNA, which is shown to be overrepresented in chromosomal translocation breakpoints in human cancer (67). Enrichment of DSBs at the highly stable DNA secondary structure regions shown in our study provide direct support that alternative DNA structure can contribute to the generation of chromosomal rearrangements.

DNA breaks mediated by DNA secondary structure have been indirectly observed during DNA replication, transcription, and DNA repair. Multiple stem-loop structures at fragile site regions have been shown to stall replication fork progress, which causes fork collapse and/or fork reversal, and both could lead to DNA breaks (60,68). Several



**Figure 7.** Evidence for the direct role of TOP2 in DNA secondary structure-mediated fragility. **(A)** Top2-1 yeast, which have decreased Top2 enzyme activity (grown at 30°C), have less gross chromosomal rearrangements with the structure-rich AT80 insert than the wild-type cells with the same insert as determined using an unpaired *t*-test ( $***P < 0.01$ ). **(B)** Knockout of TOP2B in both asynchronous and G1 synchronized RPE-1 cells results in a significantly decreased profile of cleavage complex-associated breaks at TOP2 active regions ( $***P \cong 0$ , two-sided Wilcoxon signed-rank test). Break profiles for TOP2 cleavage complex associated breaks in 65 989 identified TOP2 cleavage complex peaks for wild-type (black) and TOP2B<sup>-/-</sup> (red) cells grown asynchronously or synchronized in G1. **(C)** The DNA secondary structure forming propensity for the 65,989 peaks  $\pm$  500 bp was calculated using 300-nt windows and 1-nt steps, and the average energy for each peak was calculated at the peak  $\pm$  100 bp. All peaks were binned into twenty groups based on the difference in breaks between WT and TOP2B<sup>-/-</sup>, with the greatest loss of breaks after TOP2B knockout being coded as the strongest activity sites. The median value from each binned group was plotted against the strength, and a strong correlation between site strength and secondary structure propensity is observed (Pearson's correlation  $r = 0.908$ ,  $P = 3.19 \times 10^{-8}$ ). **(D)** Four representative bins were then plotted by median free energy of folding across the entire region. The most sensitive sites for TOP2 activity (top 5%, blue line) have more highly stable secondary structure forming ability than other TOP2 active regions, with the most favorable structure forming energies at the TOP2 cleavage complex peak summits. Additionally, the median free energy of formation for all regions is lower than the genome median ( $\Delta G = -27.9$ , black dashed line).

proteins known to recognize DNA secondary structure are involved in fragile site stability, including DNA helicases, BLM, RECQ1 and WRN (69–73), structure-specific endonucleases, CtIP, ERCC1 and MUS81 (74–79), and an exonuclease, SNM1B (80). Junction-resolving enzymes, XPF and XPG are shown to cleave H-DNA in a replication-independent manner (67). Our results support these observations, in which DNA secondary structures recognized

by these structure-specific nucleases can be promoted to form by the change in DNA superhelicity from etoposide treatment (81). Not mutually excluding, our study also suggests that TOP2 could recognize and cleave DNA at sites of secondary structure formation. *In vitro* studies have demonstrated that TOP2 recognizes and preferentially cleaves DNA hairpins one nucleotide from the 3'-base of the stem, rather than using sequence specificity (33). Human



TOP2A recognizes and cleaves within the single-stranded DNA loop region of the hairpin structure formed within alpha satellite DNA (36), and human TOP2B binds and cleaves four-way junction DNA structure (34). Site-specific cleavage by TOP2 at centromeric DNA with dyad symmetries (potential to form hairpins and four-way junctions) is found in yeast, fruit fly, chicken and human (35,37). Moreover, mismatched bases which are often present in multiple stem-loop structures, when in the proximity of TOP2 cleavage sites can greatly stimulate TOP2 cleavage activity and hinder DNA end re-ligation (38,39). We have shown that the DSB frequency at intron 11 of the *RET* gene decreased with the merbarone treatment (not increased as with etoposide treatment) (82). Merbarone is a TOP2 inhibitor and prevents its scission activity, not the ligation activity. Recently, we analyzed a set of RNA Polymerase II pausing sites ( $n = 13\ 901$ ) response to etoposide treatment, and found that the TOP2 binding sites displayed a very strong overlap with the DSB sites that we mapped and with the predicted stable secondary structures (83). The DNA structure around these pausing sites and the *RET* region are multiple stem-loop structures and do not contain DNA repeats. Both observations suggest that stable multiple stem-loop structures from non-repeating sequences, which contain more mismatched bases and single-stranded regions, could be the targeted sites for TOP2, especially under negative supercoiling. On the other hand, stable structures generated by long repeats such as perfect palindrome sequences have an absence of TOP2 cleavage (Kirill Lobachev, personal communication). Further studies to identify and investigate finer structural features or other parameters that TOP2 and other structure-specific nucleases recognize will advance our understanding of DNA secondary structure-mediated DSBs.

Two major TOP2 activity sites, the TSS regions and CTCF-binding sites are enriched in endogenous DSBs and stable DNA secondary structures, and the enrichment is tightly linked to gene expression levels and CTCF binding strength, respectively. TOP2 is known to position at the base of topologically associating domains through the interactions with cohesin and CTCF (84). Recently, Canela *et al.* showed that chromosome loop anchors are prone to TOP2-mediated DNA breaks (85) and the conversion to persistent DSBs is enhanced in actively transcribed regions (13), and Gothe *et al.* (14) demonstrated a dependence of DNA fragility on both CTCF binding and relative direction of active transcription, both of which support our observations (Figure 5). The genome-wide break analyses of Yan *et al.* (86) and Lensing *et al.* (48) observed a similar transcription dependence for DSBs at TSS regions. Active transcription generates single-stranded DNA, and we have shown that non-template strands at the TSS regions are significantly and energetically favorable to form alternative DNA secondary structure (8). Supercoiling and/or strand unwinding at TSS regions during active transcription and at CTCF-binding sites upon protein binding can promote DNA secondary structure formation at sequences with such potential to provide target sites for TOP2. Therefore, we proposed that DNA regions having potential to form highly stable secondary structures can signal excessive supercoiling, and the presence of these structures then activates the action of

TOP2 to remove the excess. Most of the DSBs generated by TOP2 are short-lived because of the re-ligation activity of these enzymes, and thus are benign to cells, supporting a physiological relevance of this model.

On the other hand, TOP2-mediated breaks are also often associated with pathological damage due to the use of topoisomerase inhibitors as anticancer drugs (17,87). Recent work has identified that TOP2 damage near CTCF is associated with translocation partners, particularly known in therapy-related acute myeloid leukemias, and that breakage and translocation events are increased with etoposide treatments (14). To further investigate the structural features that TOP2 recognizes, and to identify factors which influence such features, will change how we evaluate the effect of many anti-TOP2 chemotherapeutic agents. In addition, environmental agents like benzene (88) and phytochemicals in food and natural products (89) are known to generate TOP2-induced breaks, and identifying regions prone to these stresses will have a broad impact on our understanding of human genome instability.

## DATA AVAILABILITY

DSB mapping data can be accessed at the NCBI Sequence Read Archive (SRA), accession numbers: PRJNA542485 for endogenous break mapping in nuclei of NPC, HeLa and GM13069 cells; and PRJNA497476 for break mapping using genomic DNA purified from etoposide-treated GM13069 and untreated control cells (49). The publicly available datasets used in this study are listed in Supplementary Table S3.

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Nadine Michel and Michael McConnell for providing NPC9429 cells.

## FUNDING

National Institutes of Health [RO1GM101192 to Y.H.W., RO1GM129119 to K.S.L.]; National Cancer Institute [RO1CA113863 to Y.H.W.]; H.M.R. was supported by the Cellular and Molecular Biology Predoctoral Training Grant from the NIH [T32GM008136]. Funding for open access charge: National Institutes of Health.

*Conflict of interest statement.* None declared.

## REFERENCES

- Xie, K.T., Wang, G., Thompson, A.C., Wucherpfennig, J.I., Reimchen, T.E., MacColl, A.D.C., Schluter, D., Bell, M.A., Vasquez, K.M. and Kingsley, D.M. (2019) DNA fragility in the parallel evolution of pelvic reduction in stickleback fish. *Science*, **363**, 81–84.
- Tubbs, A. and Nussenzweig, A. (2017) Endogenous DNA damage as a source of genomic instability in cancer. *Cell*, **168**, 644–656.
- Ceccaldi, R., Rondinelli, B. and D'Andrea, A.D. (2016) Repair pathway choices and consequences at the Double-Strand break. *Trends Cell Biol.*, **26**, 52–64.

4. Wang, G. and Vasquez, K.M. (2017) Effects of replication and transcription on DNA structure-related genetic instability. *Genes (Basel)*, **8**, 17.
5. Irony-Tur Sinai, M. and Kerem, B. (2018) DNA replication stress drives fragile site instability. *Mutat. Res.*, **808**, 56–61.
6. Kaushal, S. and Freudenreich, C.H. (2019) The role of fork stalling and DNA structures in causing chromosome fragility. *Genes Chromosomes Cancer*, **58**, 270–283.
7. Dillon, L.W., Pierce, L.C., Ng, M.C. and Wang, Y.H. (2013) Role of DNA secondary structures in fragile site breakage along human chromosome 10. *Hum. Mol. Genet.*, **22**, 1443–1456.
8. Szlachta, K., Thys, R.G., Atkin, N.D., Pierce, L.C.T., Bekiranov, S. and Wang, Y.H. (2018) Alternative DNA secondary structure formation affects RNA polymerase II promoter-proximal pausing in human. *Genome Biol.*, **19**, 89.
9. Durkin, S.G. and Glover, T.W. (2007) Chromosome fragile sites. *Annu. Rev. Genet.*, **41**, 169–192.
10. Madabhushi, R., Gao, F., Pfenning, A.R., Pan, L., Yamakawa, S., Seo, J., Rueda, R., Phan, T.X., Yamakawa, H., Pao, P.C. *et al.* (2015) Activity-induced DNA breaks govern the expression of neuronal Early-Response genes. *Cell*, **161**, 1592–1605.
11. Schwer, B., Wei, P.C., Chang, A.N., Kao, J., Du, Z., Meyers, R.M. and Alt, F.W. (2016) Transcription-associated processes cause DNA double-strand breaks and translocations in neural stem/progenitor cells. *Proc. Natl Acad. Sci. U.S.A.*, **113**, 2258–2263.
12. D'Alessandro, G. and d'Adda di Fagnana, F. (2017) Transcription and DNA Damage: Holding hands or crossing swords? *J. Mol. Biol.*, **429**, 3215–3229.
13. Canela, A., Maman, Y., Huang, S.N., Wutz, G., Tang, W., Zagnoli-Vieira, G., Callen, E., Wong, N., Day, A., Peters, J.M. *et al.* (2019) Topoisomerase II-Induced chromosome breakage and translocation is determined by chromosome architecture and transcriptional activity. *Mol. Cell*, **75**, 252–266.
14. Gothe, H.J., Bouwman, B.A.M., Gusmao, E.G., Piccinno, R., Petrosino, G., Sayols, S., Drechsel, O., Minneker, V., Josipovic, N., Mizi, A. *et al.* (2019) Spatial chromosome folding and active transcription drive DNA fragility and formation of oncogenic MLL translocations. *Mol. Cell*, **75**, 267–283.
15. Gittens, W.H., Johnson, D.J., Allison, R.M., Cooper, T.J., Thomas, H. and Neale, M.J. (2019) A nucleotide resolution map of Top2-linked DNA breaks in the yeast and human genome. *Nat. Commun.*, **10**, 4846.
16. Vos, S.M., Tretter, E.M., Schmidt, B.H. and Berger, J.M. (2011) All tangled up: how cells direct, manage and exploit topoisomerase function. *Nat. Rev. Mol. Cell Biol.*, **12**, 827–841.
17. Pommier, Y., Sun, Y., Huang, S.N. and Nitiss, J.L. (2016) Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. *Nat. Rev. Mol. Cell Biol.*, **17**, 703–721.
18. Hoa, N.N., Shimizu, T., Zhou, Z.W., Wang, Z.Q., Deshpande, R.A., Paull, T.T., Akter, S., Tsuda, M., Furuta, R., Tsutsui, K. *et al.* (2016) Mre11 is essential for the removal of lethal topoisomerase 2 covalent cleavage complexes. *Mol. Cell*, **64**, 580–592.
19. Cortes Ledesma, F., El Khamisy, S.F., Zuma, M.C., Osborn, K. and Caldecott, K.W. (2009) A human 5'-tyrosyl DNA phosphodiesterase that repairs topoisomerase-mediated DNA damage. *Nature*, **461**, 674–678.
20. Zeng, Z., Cortes-Ledesma, F., El Khamisy, S.F. and Caldecott, K.W. (2011) TDP2/TTRAP is the major 5'-tyrosyl DNA phosphodiesterase activity in vertebrate cells and is critical for cellular resistance to topoisomerase II-induced DNA damage. *J. Biol. Chem.*, **286**, 403–409.
21. Gomez-Herreros, F., Romero-Granados, R., Zeng, Z., Alvarez-Quilon, A., Quintero, C., Ju, L., Umans, L., Vermeire, L., Huylebroeck, D., Caldecott, K.W. *et al.* (2013) TDP2-dependent non-homologous end-joining protects against topoisomerase II-induced DNA breaks and genome instability in cells and in vivo. *PLoS Genet.*, **9**, e1003226.
22. Gao, R., Schellenberg, M.J., Huang, S.Y., Abdelmalak, M., Marchand, C., Nitiss, K.C., Nitiss, J.L., Williams, R.S. and Pommier, Y. (2014) Proteolytic degradation of topoisomerase II (Top2) enables the processing of Top2.DNA and Top2.RNA covalent complexes by tyrosyl-DNA-phosphodiesterase 2 (TDP2). *J. Biol. Chem.*, **289**, 17960–17969.
23. Gomez-Herreros, F., Zagnoli-Vieira, G., Ntai, I., Martinez-Macias, M.I., Anderson, R.M., Herrero-Ruiz, A. and Caldecott, K.W. (2017) TDP2 suppresses chromosomal translocations induced by DNA topoisomerase II during gene transcription. *Nat. Commun.*, **8**, 233.
24. Sciascia, N., Wu, W., Zong, D., Sun, Y., Wong, N., John, S., Wangsa, D., Ried, T., Bunting, S.F., Pommier, Y. *et al.* (2020) Suppressing proteasome mediated processing of topoisomerase II DNA-protein complexes preserves genome integrity. *Elife*, **9**, e53447.
25. Quennet, V., Beucher, A., Barton, O., Takeda, S. and Lobrich, M. (2011) CtIP and MRN promote non-homologous end-joining of etoposide-induced DNA double-strand breaks in G1. *Nucleic Acids Res.*, **39**, 2144–2152.
26. Lee, K.C., Padget, K., Curtis, H., Cowell, I.G., Moiani, D., Sondka, Z., Morris, N.J., Jackson, G.H., Cockell, S.J., Tainer, J.A. *et al.* (2012) MRE11 facilitates the removal of human topoisomerase II complexes from genomic DNA. *Biol. Open*, **1**, 863–873.
27. Aparicio, T., Baer, R., Gottesman, M. and Gautier, J. (2016) MRN, CtIP, and BRCA1 mediate repair of topoisomerase II-DNA adducts. *J. Cell Biol.*, **212**, 399–408.
28. Deshpande, R.A., Lee, J.H., Arora, S. and Paull, T.T. (2016) Nbs1 converts the human Mre11/Rad50 nuclease complex into an endo/exonuclease machine specific for Protein-DNA adducts. *Mol. Cell*, **64**, 593–606.
29. Anand, R., Jasrotia, A., Bundschuh, D., Howard, S.M., Ranjha, L., Stucki, M. and Cejka, P. (2019) NBS1 promotes the endonuclease activity of the MRE11-RAD50 complex by sensing CtIP phosphorylation. *EMBO J.*, **38**, e101005.
30. Capranico, G., Tinelli, S., Zunino, F., Kohn, K.W. and Pommier, Y. (1993) Effects of base mutations on topoisomerase II DNA cleavage stimulated by mAMSAs in short DNA oligomers. *Biochemistry*, **32**, 145–152.
31. Pourquier, P., Pilon, A.A., Kohlhagen, G., Mazumder, A., Sharma, A. and Pommier, Y. (1997) Trapping of mammalian topoisomerase I and recombinations induced by damaged DNA containing nicks or gaps. Importance of DNA end phosphorylation and camptothecin effects. *J. Biol. Chem.*, **272**, 26441–26447.
32. Been, M.D. and Champoux, J.J. (1984) Breakage of single-stranded DNA by eukaryotic type I topoisomerase occurs only at regions with the potential for base-pairing. *J. Mol. Biol.*, **180**, 515–531.
33. Froelich-Ammon, S.J., Gale, K.C. and Osheroff, N. (1994) Site-specific cleavage of a DNA hairpin by topoisomerase II. DNA secondary structure as a determinant of enzyme recognition/cleavage. *J. Biol. Chem.*, **269**, 7719–7725.
34. West, K.L. and Austin, C.A. (1999) Human DNA topoisomerase II $\beta$  binds and cleaves four-way junction DNA in vitro. *Nucleic Acids Res.*, **27**, 984–992.
35. Spence, J.M., Fournier, R.E., Oshimura, M., Regnier, V. and Farr, C.J. (2005) Topoisomerase II cleavage activity within the human D11Z1 and DXZ1 alpha-satellite arrays. *Chromosome Res.*, **13**, 637–648.
36. Jonstrup, A.T., Thomsen, T., Wang, Y., Knudsen, B.R., Koch, J. and Andersen, A.H. (2008) Hairpin structures formed by alpha satellite DNA of human centromeres are cleaved by human topoisomerase II $\alpha$ . *Nucleic Acids Res.*, **36**, 6165–6174.
37. Mills, W.E., Spence, J.M., Fukagawa, T. and Farr, C.J. (2018) Site-specific cleavage by topoisomerase 2: a mark of the core centromere. *Int. J. Mol. Sci.*, **19**, 534.
38. Bigioni, M., Zunino, F., Tinelli, S., Austin, C.A., Willmore, E. and Capranico, G. (1996) Position-specific effects of base mismatch on mammalian topoisomerase II DNA cleaving activity. *Biochemistry*, **35**, 153–159.
39. Kingma, P.S. and Osheroff, N. (1998) The response of eukaryotic topoisomerases to DNA damage. *Biochim. Biophys. Acta*, **1400**, 223–232.
40. Lorenz, R., Bernhart, S.H., Honer Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P.F. and Hofacker, I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26.
41. Michel, N., Majumdar, U.B., Lannigan, J. and McConnell, M.J. (2019) Imaging flow cytometry quantifies neural genome dynamics. *Cytometry A*, **95**, 825–835.
42. Narayanan, V., Mieczkowski, P.A., Kim, H.M., Petes, T.D. and Lobachev, K.S. (2006) The pattern of gene amplification is determined by the chromosomal location of hairpin-capped breaks. *Cell*, **125**, 1283–1296.

43. Kim, H.M., Narayanan, V., Mieczkowski, P.A., Petes, T.D., Krasilnikova, M.M., Mirkin, S.M. and Lobachev, K.S. (2008) Chromosome fragility at GAA tracts in yeast depends on repeat orientation and requires mismatch repair. *EMBO J.*, **27**, 2896–2906.
44. Zhang, Y., Saini, N., Sheng, Z. and Lobachev, K.S. (2013) Genome-wide screen reveals replication pathway for quasi-palindrome fragility dependent on homologous recombination. *PLoS Genet.*, **9**, e1003979.
45. Storic, F., Lewis, L.K. and Resnick, M.A. (2001) In vivo site-directed mutagenesis using oligonucleotides. *Nat. Biotechnol.*, **19**, 773–776.
46. Nitiss, J.L., Liu, Y.X. and Hsiung, Y. (1993) A temperature sensitive topoisomerase II allele confers temperature dependent drug resistance on amacrine and etoposide: a genetic system for determining the targets of topoisomerase II inhibitors. *Cancer Res.*, **53**, 89–93.
47. Telenti, A., Pierce, L.C., Biggs, W.H., di Iulio, J., Wong, E.H., Fabani, M.M., Kirkness, E.F., Moustafa, A., Shah, N., Xie, C. *et al.* (2016) Deep sequencing of 10,000 human genomes. *Proc. Natl Acad. Sci. U.S.A.*, **113**, 11901–11906.
48. Lensing, S.V., Marsico, G., Hansel-Hertsch, R., Lam, E.Y., Tannahill, D. and Balasubramanian, S. (2016) DSBCapture: in situ capture and sequencing of DNA breaks. *Nat. Methods*, **13**, 855–857.
49. Szlachta, K., Raimer, H.M., Comeau, L.D. and Wang, Y.H. (2020) CNCC: an analysis tool to determine genome-wide DNA break end structure at single-nucleotide resolution. *BMC Genomics*, **21**, 25.
50. Kouzine, F., Wojtowicz, D., Baranello, L., Yamane, A., Nelson, S., Resch, W., Kieffer-Kwon, K.R., Benham, C.J., Casellas, R., Przytycka, T.M. *et al.* (2017) Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. *Cell Syst.*, **4**, 344–356.
51. Tilgner, H., Grubert, F., Sharon, D. and Snyder, M.P. (2014) Defining a personal, allele-specific, and single-molecule long-read transcriptome. *Proc. Natl Acad. Sci. U.S.A.*, **111**, 9869–9874.
52. Kim, D., Langmead, B. and Salzberg, S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
53. Perlea, M., Perlea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T. and Salzberg, S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.
54. Consortium, E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
55. Dellino, G.I., Palluzzi, F., Chiariello, A.M., Piccioni, R., Bianco, S., Furia, L., De Conti, G., Bouwman, B.A.M., Melloni, G., Guido, D. *et al.* (2019) Release of paused RNA polymerase II at specific loci favors DNA double-strand-break formation and promotes cancer translocations. *Nat. Genet.*, **51**, 1011–1023.
56. Sarafidou, T., Kahl, C., Martinez-Garay, I., Mangelsdorf, M., Gesk, S., Baker, E., Kokkinaki, M., Talley, P., Maltby, E.L., French, L. *et al.* (2004) Folate-sensitive fragile site FRA10A is due to an expansion of a CGG repeat in a novel gene, FRA10AC1, encoding a nuclear protein. *Genomics*, **84**, 69–81.
57. Kouzine, F., Wojtowicz, D., Yamane, A., Resch, W., Kieffer-Kwon, K.R., Bandle, R., Nelson, S., Nakahashi, H., Awasthi, P., Feigenbaum, L. *et al.* (2013) Global regulation of promoter melting in naive lymphocytes. *Cell*, **153**, 988–999.
58. Hansel-Hertsch, R., Beraldi, D., Lensing, S.V., Marsico, G., Zyner, K., Parry, A., Di Antonio, M., Pike, J., Kimura, H., Narita, M. *et al.* (2016) G-quadruplex structures mark human regulatory chromatin. *Nat. Genet.*, **48**, 1267–1272.
59. Pearson, C.E., Tam, M., Wang, Y.H., Montgomery, S.E., Dar, A.C., Cleary, J.D. and Nichol, K. (2002) Slipped-strand DNAs formed by long (CAG)<sup>n</sup>(CTG)<sup>m</sup> repeats: slipped-out repeats and slip-out junctions. *Nucleic Acids Res.*, **30**, 4534–4547.
60. Burrow, A.A., Marullo, A., Holder, L.R. and Wang, Y.H. (2010) Secondary structure formation and DNA instability at fragile site FRA16B. *Nucleic Acids Res.*, **38**, 2865–2877.
61. Irony-Tur Sinai, M., Salamon, A., Stanleigh, N., Goldberg, T., Weiss, A., Wang, Y.H. and Kerem, B. (2019) AT-dinucleotide rich sequences drive fragile site formation. *Nucleic Acids Res.*, **47**, 9685–9695.
62. Le, H., Singh, S., Shih, S.J., Du, N., Schnyder, S., Lored, G.A., Bien, C., Michaelis, L., Toor, A., Diaz, M.O. *et al.* (2009) Rearrangements of the MLL gene are influenced by DNA secondary structure, potentially mediated by topoisomerase II binding. *Genes Chromosomes Cancer*, **48**, 806–815.
63. Meyer, C., Hofmann, J., Burmeister, T., Groger, D., Park, T.S., Emerenciano, M., Pombo de Oliveira, M., Renneville, A., Villarese, P., Macintyre, E. *et al.* (2013) The MLL recombinome of acute leukemias in 2013. *Leukemia*, **27**, 2165–2176.
64. Pendleton, M., Lindsey, R.H. Jr, Felix, C.A., Grimwade, D. and Osheroff, N. (2014) Topoisomerase II and leukemia. *Ann. N. Y. Acad. Sci.*, **1310**, 98–110.
65. Inagaki, H., Kato, T., Tsutsumi, M., Ouchi, Y., Ohye, T. and Kurahashi, H. (2016) Palindrome-Mediated Translocations in Humans: A New Mechanistic Model for Gross Chromosomal Rearrangements. *Front Genet.*, **7**, 125.
66. Kato, T., Franconi, C.P., Sheridan, M.B., Hacker, A.M., Inagakai, H., Glover, T.W., Arlt, M.F., Drabkin, H.A., Gemmill, R.M., Kurahashi, H. *et al.* (2014) Analysis of the t(3;8) of hereditary renal cell carcinoma: a palindrome-mediated translocation. *Cancer Genet.*, **207**, 133–140.
67. Zhao, J., Wang, G., Del Mundo, I.M., McKinney, J.A., Lu, X., Bacolla, A., Boulware, S.B., Zhang, C., Zhang, H., Ren, P. *et al.* (2018) Distinct mechanisms of Nuclease-Directed DNA-Structure-Induced genetic instability in cancer genomes. *Cell Rep.*, **22**, 1200–1210.
68. Zhang, H. and Freudenreich, C.H. (2007) An AT-rich sequence in human common fragile site FRA16D causes fork stalling and chromosome breakage in *S. cerevisiae*. *Mol. Cell*, **27**, 367–379.
69. Pelliccioli, A. and Muzi-Falconi, M. (2013) A blooming resolvase at chromosomal fragile sites. *Nat. Cell Biol.*, **15**, 883–885.
70. Pichierri, P., Rosselli, F. and Franchitto, A. (2003) Werner's syndrome protein is phosphorylated in an ATR/ATM-dependent manner following replication arrest and DNA damage induced during the S phase of the cell cycle. *Oncogene*, **22**, 1491–1500.
71. Bacolla, A., Wang, G., Jain, A., Chuzhanova, N.A., Cer, R.Z., Collins, J.R., Cooper, D.N., Bohr, V.A. and Vasquez, K.M. (2011) Non-B DNA-forming sequences and WRN deficiency independently increase the frequency of base substitution in human cells. *J. Biol. Chem.*, **286**, 10017–10026.
72. Kamath-Loeb, A.S., Shen, J.C., Schmitt, M.W. and Loeb, L.A. (2012) The Werner syndrome exonuclease facilitates DNA degradation and high fidelity DNA polymerization by human DNA polymerase delta. *J. Biol. Chem.*, **287**, 12480–12490.
73. Lu, X., Parvathaneni, S., Hara, T., Lal, A. and Sharma, S. (2013) Replication stress induces specific enrichment of RECQ1 at common fragile sites FRA3B and FRA16D. *Mol. Cancer*, **12**, 29.
74. Naim, V., Wilhelm, T., Debatisse, M. and Rosselli, F. (2013) ERCC1 and MUS81-EME1 promote sister chromatid separation by processing late replication intermediates at common fragile sites during mitosis. *Nat. Cell Biol.*, **15**, 1008–1015.
75. Ying, S., Minocherhomji, S., Chan, K.L., Palmal-Pallag, T., Chu, W.K., Wass, T., Mankouri, H.W., Liu, Y. and Hickson, I.D. (2013) MUS81 promotes common fragile site expression. *Nat. Cell Biol.*, **15**, 1001–1007.
76. Minocherhomji, S. and Hickson, I.D. (2014) Structure-specific endonucleases: guardians of fragile site stability. *Trends Cell Biol.*, **24**, 321–327.
77. Pepe, A. and West, S.C. (2014) MUS81-EME2 promotes replication fork restart. *Cell Rep.*, **7**, 1048–1055.
78. Wang, H., Li, Y., Truong, L.N., Shi, L.Z., Hwang, P.Y., He, J., Do, J., Cho, M.J., Li, H., Negrete, A. *et al.* (2014) CtIP maintains stability at common fragile sites and inverted repeats by end resection-independent endonuclease activity. *Mol. Cell*, **56**, 1012–1021.
79. Kaushal, S., Wollmuth, C.E., Das, K., Hile, S.E., Regan, S.B., Barnes, R.P., Haouzi, A., Lee, S.M., House, N.C.M., Guyumdzhyan, M. *et al.* (2019) Sequence and nuclease requirements for breakage and healing of a structure-forming (AT)<sup>n</sup> sequence within fragile site FRA16D. *Cell Rep.*, **27**, 1151–1164.
80. Mason, J.M., Das, I., Arlt, M., Patel, N., Kraftson, S., Glover, T.W. and Sekiguchi, J.M. (2013) The SNM1B/APOLLO DNA nuclease functions in resolution of replication stress and maintenance of common fragile site stability. *Hum. Mol. Genet.*, **22**, 4901–4913.
81. Sinden, R.R. (1994) In: *DNA Structure and Function*. Academic Press, San Diego.
82. Dillon, L.W., Pierce, L.C., Lehman, C.E., Nikiforov, Y.E. and Wang, Y.H. (2013) DNA topoisomerases participate in fragility of the oncogene RET. *PLoS One*, **8**, e75741.
83. Singh, S., Szlachta, K., Manukyan, A., Raimer, H.M., Dinda, M., Bekiranov, S. and Wang, Y.H. (2020) Pausing sites of RNA



- polymerase II on actively transcribed genes are enriched with DNA double-stranded breaks. *J. Biol. Chem.*, **295**, 3990–4000.
84. Uuskula-Reimand, L., Hou, H., Samavarchi-Tehrani, P., Rudan, M. V., Liang, M., Medina-Rivera, A., Mohammed, H., Schmidt, D., Schwalie, P., Young, E. J. *et al.* (2016) Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders. *Genome Biol.*, **17**, 182.
85. Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., Kieffer-Kwon, K. R., Pekowska, A., Zhang, H., Rao, S. S. P. *et al.* (2017) Genome organization drives chromosome fragility. *Cell*, **170**, 507–521.
86. Yan, W. X., Mirzazadeh, R., Garnerone, S., Scott, D., Schneider, M. W., Kallas, T., Custodio, J., Wernersson, E., Li, Y., Gao, L. *et al.* (2017) BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nat. Commun.*, **8**, 15058.
87. Nitiss, J. L. (2009) Targeting DNA topoisomerase II in cancer chemotherapy. *Nat. Rev. Cancer*, **9**, 338–350.
88. Son, M. Y., Deng, C. X., Hoeijmakers, J. H., Rebel, V. I. and Hasty, P. (2016) A mechanism for 1,4-benzoquinone-induced genotoxicity. *Oncotarget*, **7**, 46433–46447.
89. Ketron, A. C. and Osheroff, N. (2014) Phytochemicals as anticancer and chemopreventive topoisomerase II poisons. *Phytochem. Rev.*, **13**, 19–35.