Published in final edited form as:
Cell Rep. 2021 July 13; 36(2): 109387. doi:10.1016/j.celrep.2021.109387.

Mechanistic basis for chromosomal translocations at the *E2A* gene and its broader relevance to human B cell malignancies

Di Liu[1], **Yong-Hwee Eddie Loh**[2], **Chih-Lin Hsieh**[3], **Michael R. Lieber**[1,4,*]

[1]Departments of Pathology, Biochemistry & Molecular Biology, and Molecular Microbiology & Immunology, and Section of Molecular & Computational Biology (Department of Biological Sciences), USC Norris Comprehensive Cancer Center, University of Southern California and USC Keck School of Medicine, Los Angeles, CA, USA

[2]USC Libraries Bioinformatics Services, University of Southern California and USC Keck School of Medicine, Los Angeles, CA, USA

[3]Department of Urology, USC Norris Comprehensive Cancer Center, University of Southern California and USC Keck School of Medicine, Los Angeles, CA, USA

[4]Lead contact

SUMMARY

Analysis of translocation breakpoints in human B cell malignancies reveals that DNA double-strand breaks at oncogenes most frequently occur at CpG sites located within 20–600 bp fragile zones and depend on activation-induced deaminase (AID). AID requires single-stranded DNA (ssDNA) to act, but it has been unclear why or how this region transiently acquires a ssDNA state. Here, we demonstrate the ssDNA state in the 23 bp E2A fragile zone using several methods, including native bisulfite DNA structural analysis in live human pre-B cells. AID deamination within the E2A fragile zone does not require but is increased upon transcription. High C-string density, nascent RNA tails, and direct DNA sequence repeats prolong the ssDNA state of the E2A fragile zone and increase AID deamination at overlapping AID hotspots that contain the CpG sites at which breaks occur in patients. These features provide key insights into lymphoid fragile zones generally.

Graphical abstract

[*]Correspondence: lieber@usc.edu.

AUTHOR CONTRIBUTIONS

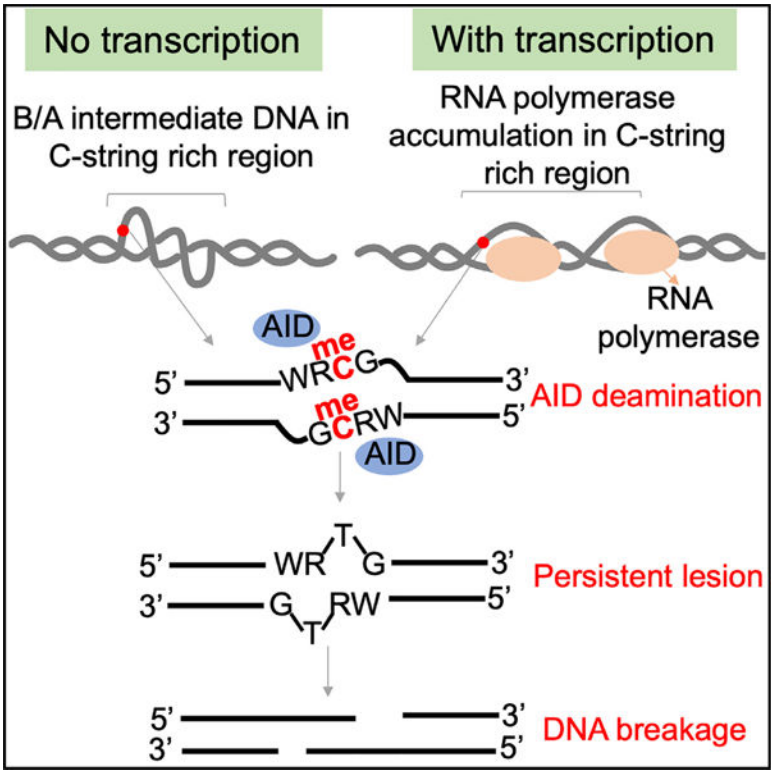
Conceptualization, methodology, and validation, D.L., C.-L.H., and M.R.L. Formal analysis and investigation, D.L., Y.-H.E.L., C.-L.H., and M.R.L. Writing – original draft, review, and editing, D.L., C.-L.H., and M.R.L. Resources, supervision, and project administration, C.-L.H. and M.R.L. Funding acquisition, M.R.L.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.109387.

DECLARATION OF INTERESTS

The authors declare no competing interests.



In brief

In patient lymphoid malignancies, there is typically an initial chromosomal translocation that involves upregulation of an oncogenic factor. The DNA breakage at these oncogenes occurs at AID hotspots that also contain CpG sites. The AID enzyme requires single-stranded DNA to act, and here Liu et al. examine how this arises.

INTRODUCTION

More than 95% of patient B cell malignancies have distinctive reciprocal chromosomal translocations (Jaffe et al., 2001; Küppers, 2005; Lieber, 2016). A chromosomal translocation event requires a double-strand break (DSB) on each of the two chromosomes that are involved. For human B cell translocations, one of the two DNA breakage events often occurs at an immunoglobulin locus. The other break is most often at an oncogene-bearing chromosome and is typically within a narrow fragile zone ranging from 20 to 600 bp (Greisman et al., 2012; Tsai et al., 2008, 2010a, 2010b). We previously reported the remarkable consistency of breakage at CpG sites (also designated CG motifs) for the most common human chromosomal translocations (Tsai et al., 2008). This provided the first clue to the basis for the breakage in the fragile zones at human B cell oncogenes.

Up to 77% of all Bcl2 breaks, 80% of all Bcl1 locus (*CCND1*) breaks, and 88% of E2A (*TCF3*) breaks are within 8 nucleotides of CpG sites. In subsequent studies, we showed that the CG motif was highly significant for being the preferred site of activation-induced deaminase (AID) action for the genes mentioned earlier and for additional B cell oncogenes, including *CRLF2* and *MALT1*, within their <600 bp fragile zones (Cui et al., 2013; Greisman et al., 2012; Pannunzio and Lieber, 2017; Tsai et al., 2008, 2010a, 2010b). In experimental systems, we showed that AID and methylation at the CpG sites were important for the propensity to break (Cui et al., 2013; Lu et al., 2015). Because the CG motif is common in the human genome, additional factors must be involved for the highly localized DNA breakage in human fragile regions. The key question remaining is, why is AID deamination activity confined to the CpG sites within the fragile zones, but not the CpG sites outside of these 20 to 600 bp fragile zones?

AID can deaminate cytosines only in regions of single-stranded DNA (ssDNA), even if this ssDNA state is transient (Pham et al., 2003, 2019). We showed that the Bcl2 and Bcl1 fragile zones have elevated bisulfite reactivity, a feature of ssDNA, at regions of consecutive uninterrupted C nucleotides (C-strings) for a length of up to 6 nt on either strand (Raghavan et al., 2004a, 2004b; Tsai et al., 2009). Based on X-ray crystallography and NMR, C-strings are known to drive duplex DNA into a conformation that is intermediate between A-form and B-form duplex DNA (called the B/A-intermediate DNA conformation) (Dornberger et al., 1999; Ng et al., 2000; Tsai et al., 2009). C-string regions are also prone to rapid opening and closing kinetics in the course of normal thermal fluctuation (Dornberger et al., 1999).

The smallest fragile zone identified thus far in B cell malignancies is the 23 bp E2A fragile zone (Lieber, 2016; Tsai et al., 2008). This fragile zone is shorter than the duplex DNA wrapped around a single histone octamer, making it distinctive among the B cell fragile zones. Sequence analysis of all published and previously unpublished breakpoint sequences (Fischer et al., 2015; Hein et al., 2019; Hunger et al., 1992; Inaba et al., 1992; Kato et al., 2017; Paulsson et al., 2007; Tsai et al., 2008; Wiemels et al., 2002) confirms that most chromosomal breaks for the *E2A* gene occur in the 23 bp E2A fragile zone. Furthermore, CG and AID WRC (W = A or T, R = A or G) hotspot motifs are in highly significant proximity to the E2A breakpoints in the motif analysis. In this study, we focus on the larger context of the E2A fragile zone to determine why the CpG sites within it are more vulnerable to breakage than the ones outside the fragile zone. We use a highly efficient method of bisulfite chemical probing of the DNA structure to rapidly mark the E2A fragile zone and the downstream region for single-stranded character in live human pre-B cells and confirm this finding in a DNA melting study and a nuclease cutting assay. We also use a biochemically defined AID system coupled with transcription and find elevated AID deamination activity in the 23 bp E2A fragile zone, as well as its immediate downstream region. In addition, transcription and the nascent transcript attached to the RNA polymerase increase the AID deamination rate within the 23 bp E2A fragile zone and the downstream region. We show that interrupting the C-strings in the E2A fragile zone and its downstream region decreases AID activity in these regions, likely by affecting the local DNA conformation. We propose that cytosines located in the WRCG motif in regions with high C-string density are prone to deamination by AID, leading to persistent DNA lesions and then DSBs. We generalize these findings to the other human B cell fragile zones to formulate a model for the features that define fragile zones in the human genome.

RESULTS

E2A breakpoints cluster around CpG sites with extremely high significance in patients

E2A sequences are available, from translocation positive patients described previously, for breakpoint analysis and sequence motif analysis (Fischer et al., 2015; Hein et al., 2019; Hunger et al., 1992; Inaba et al., 1992; Kato et al., 2017; Paulsson et al., 2007; Tsai et al., 2008; Wiemels et al., 2002). The most common translocation partner of *E2A* is the *PBX1* gene (Figure S1A). Sixty junctional sequences from 49 patients with E2A-PBX1 translocations are available (Table S2). Fifty-nine of the 60 E2A breakpoints (98%) from E2A-PBX1 translocations are mapped to *E2A* intron 16. Forty-six of the 60 E2A breakpoints (77%), including 25 from single-breakpoint boundary analysis (the sequence from only one of the translocation partners is available) and 21 from the 11 pairs of double-breakpoint boundary analysis (sequences of both translocation partners are available), are clustered in the 23 bp E2A fragile zone (Figure 1A). The remaining breakpoint of a double-breakpoint pair is only one nucleotide outside of the 23 bp fragile zone and is treated as in the 23 bp fragile zone for analyses described later. As described in STAR Methods, calculation based on these findings indicates that this region is more than 400-fold more fragile than the surrounding DNA within the 3.3 kb intron 16.

The sequence motifs at or near each breakpoint are also informative. Thirty-one of the E2A 59 breakpoints within intron 16 are right at CpG sites. The boundaries of the initial DNA breakage site on each of the two involved chromosomes can be more clearly identified when sequences of breakpoints from the reciprocal translocation partners are available (Figure S1B) (Lieber, 2016). All eight reciprocal translocations with a single or no breakpoint at a CpG site have a CpG between the pair of identified breakpoints. This suggests that the initial E2A breakage may arise at a CpG site, with some broken ends being resected by a few nucleotides, as is typical for almost all DSB junctions repaired by nonhomologous DNA end joining (NHEJ) (Zhao et al., 2020). Statistical analysis shows that proximity of E2A breakpoints to the CG motif is highly significant (Figure 1B; $p = 8.3 \times 10^{-6}$ in U test). Besides the CG motif, the AID hotspot motifs, WRC ($p = 1.1 \times 10^{-3}$ in U test) and its related motif WGCW ($p = 1.6 \times 10^{-3}$ in U test), are statistically highly significant for their proximity to E2A breakpoints (Figure 1B) (Pham et al., 2003; Yu et al., 2004).

Another common translocation partner of *E2A* is the *HLF* gene (Figure S1A). Twelve breakpoints, four single breakpoint, and four double breakpoints, sequenced from 8 patients with E2A-HLF translocations, are available for breakpoint and motif analyses (Table S3). Seven breakpoints are mapped in intron 16, and all are in the 23 bp fragile zone (Figure 1C). Six of these seven E2A breakpoints are precisely at a CpG site in the 23 bp fragile zone ($p = 1.3 \times 10^{-3}$ in the binomial test). The proximity of E2A breakpoints in the E2A-HLF translocation to the CpG site ($p = 6.9 \times 10^{-3}$ in U test) and the AID WRC hotspot motif ($p = 2.4 \times 10^{-2}$ in U test) is statistically significant (Figure 1D).

The remarkable consistency of E2A break sites within the 23 bp fragile zone in patients with the E2A-PBX1 and E2A-HLF translocations indicates the importance of this 23 bp of DNA, on either a sequence level or a DNA structural level resulting from the sequence. It also suggests that the CG motif and AID may play critical roles in the E2A breakage process.

However, AID requires DNA to be at least single stranded transiently (Pham et al., 2003), and this fact compelled us to focus extensively on this point in our study of the mechanistic basis for this more than 400-fold increased fragility in this small region of DNA.

Identification of the transient ssDNA character in the 23 bp E2A fragile zone in human cells

We wondered whether the 23 bp E2A fragile zone manifests thermal fluctuation (breathing) that can lead to transient single-strandedness of DNA focally. Bisulfite chemical probing of the DNA structure under native (nondenaturing) conditions is an ideal high-resolution method to assess stable or transient single-stranded states, as we have previously described (Raghavan et al., 2004a, 2004b, Raghavan et al., 2006; Tsai et al., 2009; Yu et al., 2003, 2006). Cytosine conversion in the native bisulfite assay is a measure of transient single-stranded character in double-stranded DNA (dsDNA).

We treated live human pre-B cells (Nalm-6 and 697 cell lines) with ammonium bisulfite and examined the 23 bp E2A fragile zone and the surrounding regions by amplifying a 270 bp region from the nontranscribed strand (NTS) and a 276 bp region from the transcribed DNA strand (TS) for next-generation sequencing (Figure 2). Combining the sequence information from NTS and TS amplicons, the total length of the E2A region examined is 389 bp with a 157 bp overlap, which includes the 23 bp E2A fragile zone. In both live Nalm-6 and 697 pre-B cells, two major bisulfite reactive peaks are clearly higher than the baseline average of single C (isolated cytosines surrounded by G, T, or A nucleotides) conversion at sites outside of the 23 bp fragile zone (labeled peak 1 and peak 2 in Figure 2). Peak 1 observed in the native bisulfite assay is located at the edge of the 23 bp E2A fragile zone on the NTS at the sequence motif of C_5AG_4, and peak 2 is ~100 bp downstream of the 23 bp E2A fragile zone on the TS, also at a C_5 motif. Further analysis of the sequence features indicates that the bisulfite conversion reactivity correlates well with increasing length of C-strings rather than other sequence features or the base composition (Figure S2). The density of C-strings with a length of 4 and greater is 1/37 bp in the 666 bp region containing the 23 bp E2A fragile zone and its downstream region, which is two-fold higher than the C-string density of the entire *E2A* intron 16 (1/72 bp) and higher than that in the region farther downstream (1/363 bp from downstream of the 666 bp zone to the end of the intron). The high C-string density in the E2A fragile zone and the 600 bp downstream region suggests the single-stranded character may extend downstream of the 23 bp fragile zone.

Importantly, the conversion rate at the single Cs within the 23 bp E2A fragile zone and the downstream region is higher than the average single C conversion baseline, which is established based on the average conversion rate of single Cs located outside of the E2A fragile zones. Specifically, the average conversion rate at single C sites within the 23 bp fragile zone is 6.3%, which is ~2.5-fold higher than the baseline of 2.4% in the 697 cells ($p = 1.1 \times 10^{-2}$; see STAR Methods for details). Furthermore, the single (i.e., isolated) C sites within the 23 bp fragile zone in both cell types are converted at multiple consecutive single Cs on the same molecule for most molecules, indicating the existence of increased single-stranded character of more than a few bases in length within the fragile zone (Figure 2C). Multiple consecutive single C conversions are observed outside the E2A fragile zone to its downstream region in Nalm-6 cells, supporting the presence of single-stranded character in the broader region harboring the 23 bp fragile zone and its downstream region. The native bisulfite assay using live cells reveals the transient (i.e., intermittent) single-stranded character in the 23 bp E2A fragile zone and the downstream region in the *E2A* intron 16 examined, suggesting frequent thermal fluctuations. The ssDNA character of the 23 bp E2A fragile zone is underestimated because of the bisulfite resistance of methylated cytosines at CpG sites (3.4% to 16.1% at the CpG sites in 697 and Nalm-6 cells, respectively; Table S4). Importantly, we do not observe a long region of asymmetry of cytosine conversions on NTS versus TS, and this definitively rules out an R-loop at the E2A fragile zone or the downstream region (Figure 2) (Yu et al., 2003; Yu and Lieber, 2019).

To explore the influence of DNA sequence on the thermal fluctuation of the 23 bp E2A fragile zone, we next performed a DNA melting study using a 27 bp substrate that contains the 23 bp E2A fragile zone and a randomized control substrate of the same nucleotide composition and length (sequence in Table S1). The three nucleotides at each edge of the E2A and the control substrates were kept identical. The melting curves were done in three sodium phosphate concentrations ranging from 10 to 50 mM (Figure 3A). Under all three conditions, the melting temperature of the 27 bp E2A substrate is ~2°C lower than that of the control substrate (Table S5). These data show that the E2A duplex is less stably base paired than the randomized control duplex with the identical nucleotide composition, indicating that the E2A fragile zone is prone to DNA thermal fluctuation.

For the thermal melting study, long regions of duplex DNA (e.g., hundreds of base pairs) are not suitable due to the high temperature required. To capture the transient DNA breathing within the 23 bp E2A fragile zone on a longer region, a P1 nuclease assay was performed on the 150 bp E2A substrate and a 150 bp control substrate. Both of the 150 bp DNA substrates have internal radioactive labels at position 76 to avoid rapid signal loss because of P1 cutting at the DNA ends (Figure S3A). P1 digestion of the 150 bp E2A substrate results in several smaller DNA fragments, some of which are derived from the 23 bp E2A fragile zone, whereas the P1 reactions with the control sequence do not show DNA fragments at the corresponding positions (Figure S3B). P1 digestion of the substrates followed by a SacI restriction enzyme digestion allows us to determine the P1 cutting sites on the 150 bp E2A substrate more precisely. Comparison of the DNA fragments generated before and after SacI digestion identifies three P1 cutting sites in the 23 bp E2A fragile zone, as indicated by red blocks, and four sites in the downstream region, as indicated by gray blocks (Figure 3B; Figure S3C). The P1 nuclease digestion results confirm the transient single-stranded character of the 23 bp E2A fragile zone and the downstream region that was revealed by the native bisulfite assay.

AID acts preferentially on the NTS of the 23 bp E2A fragile zone and its downstream region

Our previous studies on other 20–600 bp human lymphoid fragile regions indicated the important role of AID in DNA breakage (Cui et al., 2013; Greisman et al., 2012; Lieber, 2016; Pannunzio and Lieber, 2017; Tsai et al., 2008). The statistically significant proximity of E2A breakpoints to AID hotspot motifs described earlier strongly suggests the involvement of AID in E2A breakage. We used an *in vitro* assay to identify AID deamination sites on a 570 bp duplex DNA template harboring the 23 bp fragile zone near the center. This was done with and without transcription *in vitro*, and parallel control experiments without AID are described in detail in the STAR Methods. Following the enzymatic reaction, sites of AID deamination activity in a 176 bp region, including the 23 bp fragile zone on each molecule, were examined using maximum-depth sequencing (MDS). The MDS method ensures that each original molecule in the reaction is uniquely barcoded before amplification by PCR (Jee et al., 2016), and the AID deamination activity can be identified by C to T mutation occurring on each unique molecule in the *in vitro* enzymatic assay.

Control reactions without AID show only nonspecific background mutations lower than 0.03% (average background was 0.004%) in reactions both with and without transcription (Figure S4), excluding the possibility that transcription by T7 RNA polymerase could lead to cytosine conversions and reflecting the background errors generated by PCR and sequencing in MDS. This baseline level is calculated based on the number of C to T mutations among all unique molecules (see STAR Methods for details), making it reliable. After the appropriate background is subtracted (e.g., subtraction of up to 0.03%), it is apparent that more sites, especially within the 23 bp E2A fragile zone, were deaminated by AID on the NTS in a higher fraction of the molecules when the substrate is transcribed with T7 RNA polymerase (Figure 4A). In seven AID deamination sites, five are AID hotspots, with C to T mutation rates ranging from 0.04% to 0.20% observed on the NTS in the 176 bp of E2A substrate examined. The C to T mutation rates at these seven positions in AID-treated samples with transcription are statistically significantly higher than the background levels observed for the same condition without AID treatment ($p = 1.8 \times 10^{-3}$ compared with background level; see STAR Methods for details). Four of these seven high-deamination sites are clustered in a 9 bp region within the 23 bp E2A fragile zone, with one at a CpG site that is also the major translocation breakpoint in patients. Only limited mutations were detected on the TS, indicating AID deamination activity on this strand is below the detection limit of the MDS assay. These results suggest that the 23 bp E2A fragile zone is preferred by AID when the region is transcribed.

AID deamination activity is also detected on the NTS in the 176 bp region of E2A when there is no transcription through the substrate (Figure 4B), but this occurs at a somewhat lower rate than what is observed on the substrate with transcription (borderline significant with $p = 6.4 \times 10^{-2}$). AID deamination of cytosines occurs more often on the NTS than on the TS without transcription, just as it does with transcription, and little difference is seen on the TS with or without transcription (Figure 4). Four relatively high-AID deamination sites remained without transcription, with mutation rates ranging from 0.03% to 0.11% ($p = 2.1 \times 10^{-2}$ compared with background level), and all four sites, including the second CpG site in the 23 bp E2A fragile zone, are AID deamination hotspots. As we discuss later, only methylated CpG sites have the potential to become long-lived T:G mismatch lesions after they are deaminated by AID (see Discussion). Other AID hotspots may give rise to peaks of C to U mutation (thus a U:G mismatch) but cannot give rise to a long-lived lesion.

The locations of AID deamination hotspots within the 23 bp fragile zone identified in these experiments are consistent with the P1 nuclease digestion sites and sites with single-stranded character in the native bisulfite assay described earlier (Figure 3B). The deamination activity of AID detected in the 176 bp region of the E2A substrate both with and without transcription shows that cytosines in the 23 bp E2A fragile zone are preferred targets for AID and that this region can be single stranded transiently, as required for AID substrates (Pham et al., 2003, 2016; Qiao et al., 2017). In addition, the similar AID deamination activity on substrates with and without transcription indicates that this single-stranded feature on the NTS can occur even when there are no extrinsic factors to change the duplex nature of the substrate. Because AID can only act on cytosines in regions of ssDNA, the increase of AID deamination activity in the 23 bp E2A fragile zone and the downstream region by transcription indicates that transcription increases the duration and DNA length over which single-strandedness in these regions occurs.

RNase A diminishes the AID deamination activity on E2A

RNase A specifically degrades single-stranded RNA. RNase A removal of nascent RNA transcripts from T7 RNA polymerase reduces transcription stalling (Bentin et al., 2005). Therefore, we investigated the effect of RNase A on AID deamination activity on the transcribed E2A substrate. We find that the AID deamination activity in the 176 bp region is greatly reduced when RNase A is added to the E2A transcription reaction, detecting only a single clear deamination site, which is an AID hotspot motif, on the NTS in the region downstream of the 23 bp E2A fragile zone (Figure 5). Although the mutation rate of this sole AID deamination site is 0.06%, it is far lower than the 0.20% observed when RNase A is absent from the reaction. All other mutation sites detected in conditions free of RNase A decrease to negligible levels when RNase A is present ($p = 8.8 \times 10^{-3}$) (Figures 4A and 5; Figure S5). The near elimination of AID deamination activity on the NTS of E2A suggests the loss of single-stranded character of the E2A substrate by the addition of RNase A during transcription. This may be caused by the more rapid transit of the RNA polymerase after removal of the nascent RNA tail.

C-strings within the E2A substrate are essential for targeting of AID deamination

Our previous studies have indicated the presence of B/A-intermediate DNA structures at regions with C-strings in human Bcl1 and Bcl2 fragile zones (Raghavan et al., 2004a, 2004b; Tsai et al., 2009). In the current study, we detected single-stranded character at long C-strings and relatively high AID deamination activity at the 5C-string within the 23 bp E2A fragile zone. We next tested whether the C-strings in the 23 bp E2A fragile zone and the downstream region contribute to structural fluctuations, which predispose the region to AID deamination activity.

Two mutated E2A sequences (Figure S6A) were used to test this hypothesis in the same experimental conditions as used earlier. A string of five Cs at the edge of the 23 bp E2A fragile zone on the NTS was disrupted on both mut1 and mut2 mutants, and an additional two strings of four Gs on the NTS (two strings of four Cs on the TS) on mut2 were disrupted. The C-string disruptions do not add new, or delete the previously existing, AID hotspot motifs on the mut1 and mut2 substrates. The disruption of C-strings on mut1 and mut2 allows us to investigate the contribution of these DNA motifs to AID deamination activities on E2A compared with the results of the wild-type (WT) E2A substrate.

Similar to the WT E2A substrate, the same seven high-AID deamination sites are present on the NTS of mut1, with mutation rates ranging from 0.05% to 0.16% with transcription (Figure 6A). The number of high-AID deamination sites reduces to three on the NTS of mut1 in the absence of transcription with mutation rates between 0.11% and 0.20% (Figure 6B). Comparison of the AID deamination activity on the WT E2A substrate and mut1 reveals a similar deamination pattern of AID on both strands of the substrate in the presence and in absence of transcription (no statistical difference, with a $p > 0.05$; Figure S6). These results indicate that the elimination of the 5C-string on the NTS at the boundary of the 23 bp E2A fragile zone has no detectable effect on AID deamination activity on the substrate, although disruption of the 5C-string may reduce the tendency of these five bases to be single stranded (based on the findings in the native bisulfite assay).

In contrast, a reduced number of AID deamination sites and lower mutation rates are observed on the mut2 substrate compared with the WT E2A substrate, both with and without transcription (Figures 6C and 6D; Figure S6C). Only four of the seven high-AID deamination sites observed on the NTS of the WT E2A substrate in the presence of transcription remained on the mut2 substrate with transcription (Figure 6C). Compared with the WT E2A substrate, it is apparent that the AID deamination activity decreases primarily in the 23 bp fragile zone and the downstream region on the NTS of the mut2 substrate in the presence of transcription ($p = 2.7 \times 10^{-3}$; top panel of Figure S6C). Unlike the WT or mut1 substrate, no clear AID deamination site is observed on the mut2 substrate in the absence of transcription (Figure 6D). Rather, there is decreased AID deamination activity across the entire mut2 substrate compared with the WT E2A substrate ($p = 2.1 \times 10^{-2}$; bottom panel of Figure S6C). The reduced AID deamination activity on mut2 strongly suggests that disruption of one C-string on the NTS and two C-strings on the TS markedly changes the single-stranded character and accessibility of AID to the mut2 substrate both with and without transcription.

DISCUSSION

We previously showed that human B cell translocations occur at Cs within CG and AID hotspot motifs (particularly WRCG) with a remarkable consistency, and AID is required for DNA breaks to initiate at these motifs (Cui et al., 2013; Greisman et al., 2012; Tsai et al., 2008, 2010a, 2010b). AID is known to require single-strandedness to deaminate cytosines (Pham et al., 2003, 2016, 2019; Qiao et al., 2017; Yu et al., 2004). Therefore, we wanted to understand what factors make human translocation fragile zones single stranded transiently such that they become targets of AID deamination activity. This study focuses on the fragile zone in the *E2A* gene, because its size of 23 bp is the smallest known fragile zone among spontaneous human translocations.

In this study, we find that the following factors contribute to AID deamination activity in the E2A fragile zone. First, chemical probing assays in live human B cells, DNA melting assays, and single-strand-specific nuclease digestion all demonstrate some degree of single-stranded character in the 23 bp fragile zone and 100 bp region downstream of it. Second, using purified human AID, we show that the E2A fragile zone and its downstream region are preferentially targeted for cytosine deamination by AID on the NTS. Third, transcription increases the AID targeting to this region. Fourth, we show that C-strings on either strand, which are known to cause duplex DNA to adopt a structure that is intermediate between A-form and B-form DNA, are important for AID targeting to this region. Fifth, factors that slow the movement of RNA polymerase during transcription through this region favor AID targeting.

Model for the molecular basis of the 23 bp E2A fragile zone

The fusion proteins derived from breakage within intron 16 in the E2A-PBX1 and E2A-HLF translocations cause human B cell malignancy. It is understandable that a specific intron is required within the *E2A* gene to produce an oncogenic fusion protein. But within the 3.3 kb intron 16, what accounts for the more than 400-fold preference for the 23 bp fragile zone?

First, the only two CpG sites within a substantial length (236 bp) are in the 23 bp fragile zone, are centrally located in this 236 bp region, and have a 600 bp C-string-rich zone immediately downstream (Figure 7A). These two CpG sites are methylated with frequencies ranging from 4% to 90% in human pre-B cells examined in our study and up to 100% in other hematopoietic cells (Table S4). It is clear that the DNA breaks occur predominantly at CpG sites in patients based on analysis of all available translocation breakpoints involving E2A (Figure 1; Tables S1 and S2). Although other Cs can be deaminated by AID with high frequency (Figure 4), only methylated CpG sites can give rise to the long-lived T:G mismatches (rather than U:G mismatches for unmethylated Cs) that are important for the translocation process (Pannunzio and Lieber, 2017; Pfeifer, 2006; Schmutte et al., 1995; Tsai et al., 2008; Walsh and Xu, 2006).

Second, as mentioned earlier, the two CpG sites in the E2A fragile zone are located at the beginning of a region with the highest density of C-strings on both strands in the entire 3.3 kb intron 16. We have shown previously that C-strings cause duplex DNA to favor a B/A-intermediate DNA structure *in vitro* and *in vivo* and can be a target for deamination by AID (Tsai et al., 2009). In the current study, we specifically show that the 23 bp E2A fragile zone and the immediate downstream region have a high propensity to be single stranded (Figures 2 and 3; Figures S2 and S3), and AID can act on this region of DNA even without transcription. With mutant sequences altered at one or more of the C-strings in the 23 bp fragile zone and its downstream region, we show that disruption of multiple C-strings and direct repeats flanking the fragile zone reduces AID targeting to these regions (mut2 in Figure 6 and Figure S6). These results strongly indicate that multiple C-strings in the 23 bp fragile zone and its downstream region are critical and have a direct effect on AID deamination activity in this region.

Third, we show that transcription increases the AID targeting to this region beyond the direct effect of C-strings in the DNA. Transcription separates the strands, thereby providing additional ssDNA exposure around the transcription bubble. Transcription is known to be slower through regions with C-strings (Gressel et al., 2017; Pham et al., 2019; Watts et al., 2019). Slowed transcription would increase the number of RNA polymerases in the region and increase the time and length of ssDNA exposed because of the multiple transcription bubbles. Given the high density of C-strings in the 666 bp region containing the 23 bp fragile zone and its downstream sequence, several RNA polymerases would accumulate as they progress more slowly through this region. Each RNA polymerase leaves a region of transient negative superhelical tension upstream (Liu and Wang, 1987; Sinden, 1994). With several RNA polymerases, the cumulative negative superhelical tension in the 23 bp fragile zone at the beginning of this C-string-rich region would favor unwinding of the two strands of the duplex, increasing the single-strandedness of the 23 bp fragile zone and its downstream region for the AID deamination activity, as we observe.

We show in this study that removal of the RNA tail by RNase A results in a marked reduction in AID deamination activity in the 23 bp E2A fragile zone and its downstream region (Figure 5; Figure S5). Removal of the RNA tail is known to reduce transcription stalling (Bentin et al., 2005), thus reducing the number and duration of transcription bubbles. In addition, transcription through regions with direct repeats can increase the chance of misalignment, increasing the presence of ssDNA and the chance for AID action. The slippage between the direct repeats flanking the 23 bp E2A fragile zone may also contribute to the focused E2A breakage within a narrow fragile zone, because mut2 with its direct repeat eliminated shows decreased AID deamination activity (Figure S6C).

Findings in this study provide evidence for a model for the features defining the 23 bp fragile zone within intron 16 of the *E2A* gene (Figure 7). The high density of C-strings increases thermal fluctuation, and this results in transient ssDNA in the region, making it accessible to AID even without transcription. Transcription exposes the single-stranded NTS in the transcription bubble and increases the opportunity for AID deamination of the cytosines in the bubble. In addition, the B/A-intermediate DNA conformation, the C-string density, and slippage of direct repeats flanking the fragile zone combine to slow the movement of RNA polymerases, leading to the accumulation of multiple polymerases. The accumulation of polymerases favors unwinding of the duplex DNA in the 23 bp fragile zone, which has an intrinsically lower melting temperature. The retardation of RNA polymerase movement and the unwinding of the duplex DNA in the 23 bp fragile zone allow increased opportunities for AID to deaminate its preferred targets in this region. The two AID hotspots in the 23 bp fragile zone are the only ones in the 666 bp region with high C-string density that overlap, and this is an additional factor favoring their targeting by AID (Han et al., 2011; Wei et al., 2015). The deamination of cytosines in the AID hotspot motif (WRC) results in U:G mismatches that can be efficiently repaired by uracil DNA glycosylase (UDG). However, the T:G mismatch resulting from AID deamination of the 5-methylcytosine within the WRCG motif is long lived, because thymine DNA glycosylase is in low abundance in mammalian cells and because of the inefficiency of enzymes processing a T:G mismatch (Schmutte et al., 1995; Walsh and Xu, 2006). The persistent T:G mismatch can be cut by activated Artemis or the RAG complex and become a DSB (Cui et al., 2013; Pannunzio and Lieber, 2019; Tsai et al., 2008). These steps provide the opportunity to form the oncogenic fusion protein with different partners, which leads to B cell malignancy.

Relevance to other major fragile regions in human lymphoid translocations

The E2A, Bcl1, and Bcl2 fragile regions are all localized to small zones of less than 600 bp, which is different from the common human fragile sites employed in cytogenetics upon replication poison-challenge testing. The latter are millions of base pairs and do not contribute to somatic cell translocations, because their fragility is primarily only a cell culture phenomenon (Durkin and Glover, 2007). In contrast, natural DNA breakage in the B lymphoid malignancies occurs at the pro-B and pre-B cell stages, and these break sites show clear evidence of TdT activity that is characteristic of these stages of lymphoid development (Jäger et al., 2000; Tsai et al., 2008; Welzel et al., 2001). With CG as the most significant dinucleotide motif in all fragile regions with AID-initiated lesions, a new type of breakage mechanism, the AID-type break, was proposed (Cui et al., 2013; Greisman et al., 2012; Lu et al., 2015; Tsai et al., 2008, 2010a).

The breakpoints in these 20–600 bp fragile regions have highly significant proximity to AID hotspot motifs. The expression of AID in B cells, but not T cells, explains the lineage specificity of the DNA breakage (B lineage rather than T lineage). The expression level of AID at the pro-B cell and pre-B cell stages is low but sufficient to cause rare translocation events (Cantaert et al., 2015; Han et al., 2007; Kelsoe, 2014; Kumar et al., 2013; Kuraoka et al., 2009, 2011; Mao et al., 2004; Ueda et al., 2007; Umiker et al., 2014). The CG motif and its surrounding sequences determine the sequence specificity of the DNA breakage. Bcl1 and Bcl2 fragile zones adopt a B/A-intermediate DNA structure, characterized by many different assays, including native bisulfite probing (Javadekar et al., 2018; Raghavan et al., 2004a, 2004b; Tsai et al., 2009). As the only DNA motif that can be methylated, CpG, especially the one preferred by AID in WRCG motifs in the B/A-intermediate DNA zones, delineates the fragile region. The CpG sites at Bcl1 and Bcl2 breakpoints are typically within WRCG motifs and regions rich in C-strings, as we have described here for the E2A fragile zone. Deamination of the methylated CpG at these fragile zones would lead to long-lived DNA lesions (T:G mismatches) in the same manner. Our model (RAG or activated Artemis cutting at AID-deaminated methylated CpG sites) explains the developmental stage, lymphoid lineage, and sequence specificity of the major lymphoid translocation breakpoints (Figure 7). The study here explains how the essential ssDNA required for AID action arises, and an explanation for this critical first step has been lacking up to this time.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
T4 Polynucleotide Kinase (PNK)	New England BioLabs	Cat# M0201S
T4 DNA ligase	Millipore Sigma	Cat# 10716359001
P1 nuclease	New England BioLabs	Cat# M0660S
Q5 high-fidelity DNA polymerase	New England BioLabs	Cat# M0491S
Taq DNA polymerase	M.R.Lieber lab	N/A
sodium bisulfite	Sigma	Cat# S-9000
sodium hydroxide	J.T.Baker	Cat# 3722–01
hydroquinone	Sigma	Cat#H-9003
$(NH_4)_2SO_3$●H_2O	Wako	Cat# 7783–11-1

REAGENT or RESOURCE	SOURCE	IDENTIFIER
50% NH_4HSO_3	Wako	Cat# 014–02905
Oligonucleotides		
Sequences listed in Table S1	Integrated DNA Technologies	N/A
Recombinant DNA		
Human GST-AID	Pham et al., 2019	N/A
Software and algorithms		
Matlab2018a	Mathworks	N/A
Microsoft Excel	Microsoft	N/A
Others		
Wizard® DNA Clean-Up System	Promega	Cat#A7280
HighPrep PCR system	MAGBIO	Cat# AC-60050

RESOURCE AVAILABILITY

Lead contact—Further information requests may be directed to, and will be fulfilled by, the Lead Contact, Dr. Michael R. Lieber (lieber@usc.edu).

Materials availability—This study did not generate new unique reagents.

Data and code availability

- All of the DNA sequencing is done after bisulfite treatment or after AID enzyme treatment. These were used to assess DNA structure or DNA methylation status. Bisulfite and AID alter the original sequence, and therefore, there is no native DNA sequence in the paper, and only experimentally altered sequence remains. For this reason, we have not deposited the experimentally altered sequence.
- This study did not generate original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Three human pre-B cell lines, including 697, Reh, and Nalm-6 were used in this study. The pre-B cells were cultured in RPMI 1640 medium supplemented with 10% fetal bovine serum. Cells were maintained at a density of 2.5 × 10^7/ml and usually passaged to 0.5 × 10^7/ml. Harvested cells were washed twice with 1 × PBS buffer before further processing.

METHOD DETAILS

DNA probes—All ssDNA oligonucleotides (oligos) were synthesized by IDT and sequences of these oligos are summarized in Table S1. The mut1 and mut2 substrates were synthesized by IDT as minigenes on plasmids. All oligos were purified using urea-denaturing PAGE unless otherwise specified in the table.

Native bisulfite assay for analysis of DNA structure—Sodium bisulfite is more commonly used for cytosine conversion experiments; however, the long incubation time required for native conditions can lead to the accumulation of non-specific cytosine conversions. Ammonium bisulfite, due to a higher achievable concentration, has been shown to be more efficient than sodium bisulfite and enables the completion of cytosine conversion within 15 to 30 min (Hayatsu et al., 2008; Shiraishi and Hayatsu, 2004). The shorter treatment time in the native ammonium bisulfite assay (< 30 min) compared with the sodium bisulfite assay (16 hours) reduces the non-specific conversions (data not shown).

The 10 M ammonium bisulfite mixture contained 2.08 g $NaHSO_3$, 0.67 g $(NH_4)_2SO_3$●H_2O (Wako), and 5 mL of 50% NH_4HSO_3 (Wako). After a homogeneous solution was achieved with 70°C incubation and intermittent mixing, the ammonium bisulfite mixture was determined to be pH 5.3. Between 1,000 to 50,000 cells were embedded in 25 µl of 1% agarose (agarose plug) and treated with 500 µl of 10 M ammonium bisulfite mixture for 30 mins at 37°C. After removal of the ammonium bisulfite mixture, the agarose plug was washed with 500 µl of water for 4 times, 500 µl of 0.3 M freshly prepared NaOH for 4 times, 500 ml of water for 3 times, and 500 µl of 10 mM Tris (pH 7.5) for 2 times. All washing steps were with 5 min incubation at room temperature for each except the last NaOH wash for 10 min. After the washing steps, the agarose plug was melted at 65°C and 1 µl of it was used in PCR. Strand-biased primers were designed to be very C-poor for the target strand and very C-rich for the undesired strand to amplify the bisulfite treated DNA with high specificity to target either the NTS or the TS in the reaction. Priming sites with no more than two cytosines on the targeted strand and more than 45% of the bases being cytosine on the undesired strand in the reaction were chosen. These primers allow amplification of all the molecules (both with and without C conversions at the priming sites) from the targeted strand and only the rare molecules without any conversion at the many Cs at the priming sites from the undesired strand in the PCR reaction. The Illumina adaptor sequences were added to each amplified sample by PCR (30 cycles of 94°C 30 s, 68°C 20 s, 72°C 1 min) using DL63 and DL64 for the NTS and DL68 and DL69 for the TS. After addition of the adaptors, a unique index (6 nucleotides) was added to each amplified sample through 25 cycles of amplification (94°C 30 s, 65°C 75 s) using i5 and i7 index primers. Amplified, adaptor added, and indexed samples from different bisulfite treated cells were pooled together and sequenced with Illumina MiSeq Reagent Kit v2 (2 × 250 bp).

Traditional (denatured) bisulfite assay for CpG methylation—Genomic DNA (gDNA) extracted from Nalm-6, Reh, and 697 cells was digested with EcoRI and denatured with final 0.6 M NaOH by incubation at 37°C for 15 min. Sodium bisulfite solution was freshly prepared with a final 2.5 M sodium metabisulfite and 2.5 M NaOH. The digested gDNA was treated with a final concentration of 2.3 M sodium bisulfite solution and 0.5 mM hydroquinone at 55°C for 4 hours in dark sealed with liquid wax. The bisulfite-treated gDNA was purified with the Wizard DNA clean-up resin (Promega) after removal of the liquid wax. Purified DNA was treated with a final concentration of 0.3 M NaOH at 37°C for

15 min followed by ethanol precipitation. Fully converted primers (DL76 and DL77) for the E2A NTS were used to amplify the target E2A region. PCR (30 cycles of 94°C 30 s, 68°C 20 s, 72°C 1 min) using DL76 and DL77 was carried out to add Illumina adaptor sequences to each sample followed by 25 cycles of index PCR (94°C 30 s, 65°C 75 s) to give each sample a unique 6 bp index using Illumina i7 and i5 primers. All samples were pooled together and sequenced by pair-end sequencing with Illumina MiSeq Reagent Kit v2 (2 × 250 bp).

Bisulfite sequencing data analysis—The sequence of the E2A NTS was used as the reference sequence for analyzing the native and denatured bisulfite sequencing data. The bisulfite conversions are reflected as C to T conversions for the molecules amplified from the NTS and the G to A conversions for the molecules amplified from the TS of E2A. Only the NTS strand was amplified in denatured bisulfite sequencing.

Native and denatured bisulfite sequencing data were first filtered with three criteria. First, the number of mismatches on each read is no greater than the number of cytosines on each read for native bisulfite sequencing and no greater than 70 (67 Cs on the NTS) for denatured bisulfite sequencing. Second, the number of other types of mutations other than C to T conversions (or G to A conversions) on each read is less than 2. Third, each R1 and R2 from the paired-end sequencing are matched in the overlapping region. Only reads passing all three criteria above were retained for further analysis. The identical reads were treated as PCR duplicates and only one was included in the native bisulfite analysis since the chance of two or more independent molecules having identical conversion pattern is low. In denatured bisulfite treatment, all unmethylated cytosines should be fully converted and carry the identical conversion pattern. Therefore, identical reads were treated as different molecules and kept for further analysis in denatured bisulfite sequencing. The reads were next sorted to three main categories: molecules containing C to T conversions (mainly arising from the NTS), molecules with G to A conversions (mainly arising from the TS), and molecules with an equal number of C to T and G to A conversions (due to hybrid reads in PCR or random mutations). Since strand-biased primers were used in the native bisulfite sequencing, reads arising from the TS and hybrid reads were removed from analysis of conversion on the NTS. Likewise, reads derived from the NTS and hybrid reads were excluded from TS conversion analysis. A total of 1,291 and 2,144 reads were included in the calculation of cytosine conversion rate on the NTS and TS of native bisulfite treated Nalm-6 cells, respectively. A total of 1,715 and 4,317 reads were included for the NTS and the TS for the 697 cell sample, respectively. The numbers of reads included in the denatured bisulfite analysis were 39,587 for 697 cells, 13,329 for Nalm-6 cells, and 20,933 for Reh cells.

The cytosine conversion rate at each position was calculated by dividing the number of reads with conversions at that position by the total reads. Therefore, the conversion rate of each cytosine in native bisulfite sequencing assay represents the relative DNA thermal fluctuation (‘breathing’) frequency of that cytosine in all molecules with conversions. The portion of unconverted cytosine in the denatured bisulfite sequencing assay represents the methylated percentage.

Melting curve analysis—The 27 nt oligos for the DNA melting study (DL94, DL95, DL96, and DL97 in Table S1) were purified by IDT using HPLC method, and directly used without further purification. The duplexes were obtained by annealing of oligos (DL94 and DL95 for the E2A substrate, DL96 and DL97 for the control substrate) in buffer of 10 mM sodium phosphate and 1 mM EDTA, pH 7.8 at 1:1 ratio. The melting reactions (20 μl) contained 480 ng of the 27 bp E2A or control substrate and 0.7x SYBR Green I in 1 mM EDTA and sodium phosphate (pH 7.8) at the indicated concentrations. The change in signal strength was recorded every 0.5°C starting from 60°C to 95°C (BioRad, C1000 Thermal Cycler).

P1 nuclease assay—The 150 bp duplex DNA substrates used in the P1 nuclease assay were prepared by ligation of two 75 nt oligos bridged by a 30 nt oligo, followed by annealing of the two 150 nt ligation products of complementary sequences. In detail, oligo DL79 was radiolabeled at its 5’ end using T4 polynucleotide kinase (New England Biolabs). DL78 and the radiolabeled DL79 were ligated with T4 DNA ligase (Roche) at 16°C overnight with DL80 as a bridge. The single-stranded 150 nt ligation product with an internal radioactive label was purified with 5% urea-denaturing PAGE after ligation and used as the NTS of E2A. Similarly, the TS of E2A was prepared with DL81, DL82, and DL83 (bridge oligo) but without radioactive labeling. The duplex E2A substrate was created by annealing of the internally-labeled NTS with excess TS. The control duplex was prepared in the same manner as described above using DL85 and the radioactive DL86 with DL87 as the bridge for the NTS and DL88, DL89, and DL90 (bridge oligo) for the TS. The 10 μl P1 reactions containing 200 fmol DNA substate and various amounts of P1 nuclease (0, 1, 2, 5, and 10 units) in 1 × P1 buffer (50 mM sodium acetate and 1 mM $ZnCl_2$, pH 5.5) were incubated at 37°C for 5 min and terminated with a final concentration of 5 mM EDTA. All reactions were phenol/chloroform extracted and allowed to let the residual chloroform evaporate without ethanol precipitation of the DNA. Half of each sample (5 μl out of 10 μl) was then treated with 2 units of SacI restriction enzyme in 1× CutSmart buffer at 37°C for 1 hour, followed by phenol/chloroform extraction and chloroform evaporation. DNA fragments from all reactions were resolved on a 6% urea-denaturing sequencing gel (8 M urea) after being boiled for 5 min with an equal volume of formamide. A mixture of oligos of various sizes between 30 nt to 82 nt in length were radiolabeled and used as size markers. The gel was autoradiographed using a phosphor-imager FX (Bio-Rad). The size of bands on the gel was determined based on migration distances of size markers.

AID deamination activity during T7 RNA polymerase transcription—The 570 bp substrates for wild-type (WT) E2A, mut1, and mut2 containing the T7 promoter were generated by digestion of pDL14, pDL31, and pDL32, respectively, with BlpI and XhoI at 37°C for 2 hours in 1× CutSmart buffer followed by native PAGE (5%) purification. The recovered DNA was quantified by real-time qPCR with DL127, DL128, and DL129 (45 cycles of 94°C 30 s, 65°C 1 min).

The AID reactions containing 40 fmol/μl of GST-AID, 250 μM rNTP mix, 10 mM DTT, and 2 units/μl of T7 RNA polymerase in 1 × transcription buffer (Promega), with 5 fmol/μl of the 570 bp DNA substrate added last to initiate the transcription reaction, were incubated at

37°C for 60 min. T7 RNA polymerase was omitted for the untranscribed reactions. Where indicated, 50 ng/μl of RNase A was added at the same time as other components to the reactions. The GST-AID variant of human AID with a carboxy (C) terminal hexa-His-tag is a generous gift from Dr. Myron Goodman (Pham et al., 2019). Reactions were terminated and purified using HighPrep PCR system (MAGBIO) and then digestion with HaeII restriction enzyme at 37°C for 2 hours was done followed by another purification with HighPrep PCR system. HaeII digestion of the reaction generated a 212 bp E2A substrate containing the 23 bp E2A fragile zone which was used as template in the Maximum-Depth Sequencing (MDS) library construction described below.

Maximum-depth sequencing (MDS) library construction for Illumina sequencing—The NTS and TS of each sample were amplified in separate reactions and subjected to MDS. MDS allows detection of the rare mutations by reducing the background error rate in sequencing and PCR (Jee et al., 2016).

Each DNA molecule from the NTS and TS amplification reactions was assigned a unique 24 nt barcode (24N) at the 3' end through a 1-cycle linear extension with Taq polymerase using 76 nt barcode primers (DL123 for the NTS and DL124 for the TS), which contain Illumina i7 adaptor sequence, 24N, and strand-specific primer sequence (95°C 3 min, ramping from 70°C to 55°C by 0.1°C/s, 55°C 2min, 72°C 2 min). Excess barcode primer and the single-stranded non-target strand were removed by exonuclease I treatment at 37°C for 1 hour.

Following cleanup by the HighPrep PCR system, the recovered DNA from the barcode assignment reaction was quantified by realtime qPCR with DL127, DL128, and DL129 (45 cycles of 94°C 30 s, 65°C 1 min). Approximately, 1 million molecules from each reaction were used as template in a 10-cycle linear amplification reaction by Taq polymerase with 0.1 nM i7 index primers (95°C 2 min, 10 cycles of 95°C 30 s, 65°C 30 s, 72°C 1 min) and the products were purified with the HighPrep PCR system. Multiple copies of the original molecule with a unique 24N barcode were generated from each molecule at this step, and the dU generated by AID deamination on each original template molecule is copied as T on these amplified molecules with the same unique barcode. Therefore, the sequence of molecules with the same barcode can be aligned to the reference sequence in the analyses to eliminate random mutations that occurred during PCR (not from AID deamination activity).

Two rounds of exponential PCR using Q5 high-fidelity DNA polymerase were performed to add unique i5 and i7 indexes to each reaction after the 10-cycle linear amplification. The final PCR products were purified with 5% native PAGE. The samples were pooled and sequenced with Illumina MiSeq Reagent Kit v3 (2 × 300 bp). All experiments were repeated at least twice for validation.

MDS data analysis—The known genomic E2A NTS and TS sequences were used as references for NTS samples and TS samples, respectively. The read pairs that contained inconsistent 24N barcode on R1 and R2 were removed first. The R1 and R2 molecules from each individual sample were sorted based on the i5 and i7 indexes and read pairs with identical 24N barcode that originated from the same original molecule were grouped together. Only the groups with at least two read pairs were retained for consensus sequence analysis. Within each group, any read with less than 70% match to the reference was considered a poor quality read, and its sequence was replaced with all Ns. A consensus sequence was called independently from all R1 reads and from all R2 reads in each group when a nucleotide is present in > 50% of the reads at the specific position. A ‘N nucleotide’ was called in the event that two different nucleotides were present at equal ratio for the same position. The consensus sequences (one for R1 and one for R2) of each group were then treated as a consensus read pair of a single molecule. Consensus read pairs from all the groups were filtered with the following criteria 1) the number of mismatches on each read is less than the number of cytosines; 2) each R1 and R2 from the paired reads are matched in the overlapping region to identify validated consensus read pairs with less than 53 mismatches on R1 and less than 55 mismatches on R2 (number of Cs on NTS and TS, respectively) to the reference sequence for base mutation calling.

The mutation frequency for each position was calculated by dividing the number of pairs with a certain mutation at that position by the total number of validated pairs. The total number of validated pairs involved in each treatment condition was listed in Table S6. The C to T mutation with background subtracted is presented in all figures (Figures 4, 5, and 6). For background subtraction, the C to T conversion rates in samples without AID treatment were regarded as background and subtracted from that of the AID treated sample under the same condition with the negative values treated as 0. Differences in AID effect as shown in Figures S5 and S6 were calculated by subtraction between background subtracted samples.

Sequencing quality at the ends of R1 and R2 reads was low and the base calling was unreliable; therefore, positions at the 3’ end (6 nt for the NTS samples and 1 nt for the TS samples) were excluded from the analysis. The amplicon analyzed and shown was 176 bp. The conversion rates of other types of mutations are all close to the background level and excluded from the figures for simplicity. Some reads contained unexpectedly high G to T mutations, and these reads were excluded from the analysis after confirmation that the changes are independent of AID activity and do not co-localize on molecules with C to T mutations.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis of patient breakpoints—tA total of 60 published and unpublished junctional sequences for E2A-PBX1 translocations from 49 patient and 12 junctional sequences for E2A-HLF translocations from 8 patients were used for statistical analysis in this study. The junctional sequences were summarized in Tables S2 and S3. The statistical analysis was performed as previously described (Tsai et al., 2008). The binomial statistic gives the probability that the E2A breakpoints occur at the tested DNA motif by random chance. The Student’s t test and the Mann-Whitney U-test are two different tests for proximity of the E2A breakpoints to a given motif (Tsai et al., 2008). Specifically, the distance of the breakpoints to the tested DNA motif is 0 bp if the breakage occurs within or next to the motif. For example, breakage at any positions marked by ‘|’ in |C|G| is defined as 0 bp from the CG motif. The fold change of fragility of the fragile zone is calculated as the number of patients per base pair in the 23 bp E2A fragile zone divided by that in the region outside of the fragile zone in *E2A* intron 16.

Statistical analyses in native bisulfite conversion rates and AID mutation rates —The Mann-Whitney U test was used in all analyses.

The comparison of single C conversion rates within the fragile zone to the background level in native bisulfite sequencing of 697 cells was performed between 6 single C positions within the fragile zone and 25 single C positions outside of the fragile zone. The average cytosine conversion rates from the NTS and TS of two different cells were grouped together according to the length of C-strings as in Figure S2C. The statistical analysis was performed between groups of different C-string length. The p value of 0.05 was used as the significance cutoff.

The comparisons between transcribed E2A samples in the AID biochemical assay were done on the seven positions shown in Figure 4A. The comparisons between E2A samples without transcription were performed using the mutation rates of the four positions in Figure 4B. When comparing AID treated samples with the background, the raw mutation rates without background subtraction were used. In other cases (e.g., the effect of RNase A, wt E2A with mut1, and wt E2A with mut2), the mutation rates after appropriate background subtraction were used (see MDS data analysis section in STAR Methods for details about the background subtraction).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank the USC Libraries Bioinformatics Service for assisting with data analysis. The bioinformatics software and computing resources used in the analysis are funded by the USC Office of Research and the Norris Medical Library. We thank all members of the Lieber lab and the Hsieh lab for helpful discussions. The authors are extremely grateful to Drs. Myron Goodman and Phuong Pham for providing purified human glutathione S-transferase (GST)-AID. This work was supported by NIH grants to M.R.L. (GM118009 and CA196671).

REFERENCES

Bentin T, Cherny D, Larsen HJ, and Nielsen PE (2005). Transcription arrest caused by long nascent RNA chains. Biochim. Biophys. Acta 1727, 97–105. [PubMed: 15716026]

Cantaert T, Schickel J-N, Bannock JM, Ng YS, Massad C, Oe T, Wu R, Lavoie A, Walter JE, Notarangelo LD, et al. (2015). Activation-Induced Cytidine Deaminase Expression in Human B Cell Precursors Is Essential for Central B Cell Tolerance. Immunity 43, 884–895. [PubMed: 26546282]

Cui X, Lu Z, Kurosawa A, Klemm L, Bagshaw AT, Tsai AG, Gemmell N, Müschen M, Adachi N, Hsieh CL, and Lieber MR (2013). Both CpG methylation and activation-induced deaminase are required for the fragility of the human bcl-2 major breakpoint region: implications for the timing of the breaks in the t(14;18) translocation. Mol. Cell. Biol 33, 947–957. [PubMed: 23263985]

Dornberger U, Leijon M, and Fritzsche H (1999). High base pair opening rates in tracts of GC base pairs. J. Biol. Chem 274, 6957–6962. [PubMed: 10066749]

Durkin SG, and Glover TW (2007). Chromosome fragile sites. Annu. Rev. Genet 41, 169–192. [PubMed: 17608616]

Fischer U, Forster M, Rinaldi A, Risch T, Sungalee S, Warnatz H-J, Bornhauser B, Gombert M, Kratsch C, Stütz AM, et al. (2015). Genomics and drug profiling of fatal TCF3-HLF-positive acute lymphoblastic leukemia identifies recurrent mutation patterns and therapeutic options. Nat. Genet 47, 1020–1029. [PubMed: 26214592]

Greisman HA, Lu Z, Tsai AG, Greiner TC, Yi HS, and Lieber MR (2012). IgH partner breakpoint sequences provide evidence that AID initiates t(11;14) and t(8;14) chromosomal breaks in mantle cell and Burkitt lymphomas. Blood 120, 2864–2867. [PubMed: 22915650]

Gressel S, Schwalb B, Decker TM, Qin W, Leonhardt H, Eick D, and Cramer P (2017). CDK9-dependent RNA polymerase II pausing controls transcription initiation. eLife 6, e29736. [PubMed: 28994650]

Han J-H, Akira S, Calame K, Beutler B, Selsing E, and Imanishi-Kari T (2007). Class switch recombination and somatic hypermutation in early mouse B cells are mediated by B cell and Toll-like receptors. Immunity 27, 64–75. [PubMed: 17658280]

Han L, Masani S, and Yu K (2011). Overlapping activation-induced cytidine deaminase hotspot motifs in Ig class-switch recombination. Proc. Natl. Acad. Sci. USA 108, 11584–11589. [PubMed: 21709240]

Hayatsu H, Shiraishi M, and Negishi K (2008). Bisulfite modification for analysis of DNA methylation. Curr. Protoc. Nucleic Acid Chem 33, 6.10.11–16.10.15.

Hein D, Dreisig K, Metzler M, Izraeli S, Schmiegelow K, Borkhardt A, and Fischer U (2019). The preleukemic TCF3-PBX1 gene fusion can be generated *in utero* and is present in z0.6% of healthy newborns. Blood 134, 1355–1358. [PubMed: 31434706]

Hunger SP, Ohyashiki K, Toyama K, and Cleary ML (1992). Hlf, a novel hepatic bZIP protein, shows altered DNA-binding properties following fusion to E2A in t(17;19) acute lymphoblastic leukemia. Genes Dev 6, 1608–1620. [PubMed: 1516826]

Inaba T, Roberts WM, Shapiro LH, Jolly KW, Raimondi SC, Smith SD, and Look AT (1992). Fusion of the leucine zipper gene HLF to the E2A gene in human acute B-lineage leukemia. Science 257, 531–534. [PubMed: 1386162]

Jaffe ES, Harris NL, Stein H, and Vardiman JW, eds. Volume 3 (IARC Press).

Jäger U, Böcskör S, Le T, Mitterbauer G, Bolz I, Chott A, Kneba M, Mannhalter C, and Nadel B (2000). Follicular lymphomas' BCL-2/IgH junctions contain templated nucleotide insertions: novel insights into the mechanism of t(14;18) translocation. Blood 95, 3520–3529. [PubMed: 10828038]

Javadekar SM, Yadav R, and Raghavan SC (2018). DNA structural basis for fragility at peak III of BCL2 major breakpoint region associated with t(14;18) translocation. Biochim. Biophys. Acta Gen. Subj 1862, 649–659. [PubMed: 29246583]

Jee J, Rasouly A, Shamovsky I, Akivis Y, Steinman SR, Mishra B, and Nudler E (2016). Rates and mechanisms of bacterial mutagenesis from maximum-depth sequencing. Nature 534, 693–696. [PubMed: 27338792]

Kato M, Ishimaru S, Seki M, Yoshida K, Shiraishi Y, Chiba K, Kakiuchi N, Sato Y, Ueno H, Tanaka H, et al. (2017). Long-term outcome of 6-month maintenance chemotherapy for acute lymphoblastic leukemia in children. Leukemia 31, 580–584. [PubMed: 27698447]

Kelsoe G (2014). Curiouser and curiouser: the role(s) of AID expression in self-tolerance. Eur. J. Immunol 44, 2876–2879. [PubMed: 25308427]

Kumar S, Wuerffel R, Achour I, Lajoie B, Sen R, Dekker J, Feeney AJ, and Kenter AL (2013). Flexible ordering of antibody class switch and V(D)J joining during B-cell ontogeny. Genes Dev 27, 2439–2444. [PubMed: 24240234]

Küppers R (2005). Mechanisms of B-cell lymphoma pathogenesis. Nat. Rev. Cancer 5, 251–262. [PubMed: 15803153]

Kuraoka M, Liao D, Yang K, Allgood SD, Levesque MC, Kelsoe G, and Ueda Y (2009). Activation-induced cytidine deaminase expression and activity in the absence of germinal centers: insights into hyper-IgM syndrome. J. Immunol 183, 3237–3248. [PubMed: 19667096]

Kuraoka M, Holl TM, Liao D, Womble M, Cain DW, Reynolds AE, and Kelsoe G (2011). Activation-induced cytidine deaminase mediates central tolerance in B cells. Proc. Natl. Acad. Sci. USA 108, 11560–11565. [PubMed: 21700885]

Lieber MR (2016). Mechanisms of human lymphoid chromosomal translocations. Nat. Rev. Cancer 16, 387–398. [PubMed: 27220482]

Liu LF, and Wang JC (1987). Supercoiling of the DNA template during transcription. Proc. Natl. Acad. Sci. USA 84, 7024–7027. [PubMed: 2823250]

Lu Z, Lieber MR, Tsai AG, Pardo CE, Müschen M, Kladde MP, and Hsieh C-L (2015). Human lymphoid translocation fragile zones are hypomethylated and have accessible chromatin. Mol. Cell. Biol 35, 1209–1222. [PubMed: 25624348]

Mao C, Jiang L, Melo-Jorge M, Puthenveetil M, Zhang X, Carroll MC, and Imanishi-Kari T (2004). T cell-independent somatic hypermutation in murine B cells with an immature phenotype. Immunity 20, 133–144. [PubMed: 14975236]

Ng H-L, Kopka ML, and Dickerson RE (2000). The structure of a stable intermediate in the A <–> B DNA helix transition. Proc. Natl. Acad. Sci. USA 97, 2035–2039. [PubMed: 10688897]

Pannunzio NR, and Lieber MR (2017). AID and Reactive Oxygen Species Can Induce DNA Breaks within Human Chromosomal Translocation Fragile Zones. Mol. Cell 68, 901–912. [PubMed: 29220655]

Pannunzio NR, and Lieber MR (2019). Constitutively active Artemis nuclease recognizes structures containing single-stranded DNA configurations. DNA Repair (Amst.) 83, 102676. [PubMed: 31377101]

Paulsson K, Jonson T, Ora I, Olofsson T, Panagopoulos I, and Johansson B (2007). Characterisation of genomic translocation breakpoints and identification of an alternative TCF3/PBX1 fusion transcript in t(1;19)(q23;p13)-positive acute lymphoblastic leukaemias. Br. J. Haematol 138, 196–201. [PubMed: 17593026]

Pfeifer G (2006). Mutagenesis at methylated CpG sequences. In DNA methylation: basic mechanisms, Doerfler W and Böhm P, eds. (Springer), pp. 259–281.

Pham P, Bransteitter R, Petruska J, and Goodman MF (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature 424, 103–107. [PubMed: 12819663]

Pham P, Afif SA, Shimoda M, Maeda K, Sakaguchi N, Pedersen LC, and Goodman MF (2016). Structural analysis of the activation-induced deoxycytidine deaminase required in immunoglobulin diversification. DNA Repair (Amst.) 43, 48–56. [PubMed: 27258794]

Pham P, Malik S, Mak C, Calabrese PC, Roeder RG, and Goodman MF (2019). AID-RNA polymerase II transcription-dependent deamination of IgV DNA. Nucleic Acids Res 47, 10815–10829. [PubMed: 31566237]

Qiao Q, Wang L, Meng FL, Hwang JK, Alt FW, and Wu H (2017). AID Recognizes Structured DNA for Class Switch Recombination. Mol. Cell 67, 361–373.e4. [PubMed: 28757211]

Raghavan SC, Houston S, Hegde BG, Langen R, Haworth IS, and Lieber MR (2004a). Stability and strand asymmetry in the non-B DNA structure at the bcl-2 major breakpoint region. J. Biol. Chem 279, 46213–46225. [PubMed: 15328356]

Raghavan SC, Swanson PC, Wu X, Hsieh C-L, and Lieber MR (2004b). A non-B-DNA structure at the Bcl-2 major breakpoint region is cleaved by the RAG complex. Nature 428, 88–93. [PubMed: 14999286]

Raghavan SC, Tsai A, Hsieh CL, and Lieber MR (2006). Analysis of non-B DNA structure at chromosomal sites in the mammalian genome. Methods Enzymol 409, 301–316. [PubMed: 16793408]

Schmutte C, Yang AS, Beart RW, and Jones PA (1995). Base excision repair of U:G mismatches at a mutational hotspot in the p53 gene is more efficient than base excision repair of T:G mismatches in extracts of human colon tumors. Cancer Res 55, 3742–3746. [PubMed: 7641186]

Shiraishi M, and Hayatsu H (2004). High-speed conversion of cytosine to uracil in bisulfite genomic sequencing analysis of DNA methylation. DNA Res. 11, 409–415. [PubMed: 15871463]

Sinden RR (1994). DNA structure and function (Gulf Professional Publishing).

Tsai AG, Lu H, Raghavan SC, Muschen M, Hsieh CL, and Lieber MR (2008). Human chromosomal translocations at CpG sites and a theoretical basis for their lineage and stage specificity. Cell 135, 1130–1142. [PubMed: 19070581]

Tsai AG, Engelhart AE, Hatmal MM, Houston SI, Hud NV, Haworth IS, and Lieber MR (2009). Conformational variants of duplex DNA correlated with cytosine-rich chromosomal fragile sites. J. Biol. Chem 284, 7157–7164. [PubMed: 19106104]

Tsai AG, Lu Z, and Lieber MR (2010a). The t(14;18)(q32;q21)/IGH-MALT1 translocation in MALT lymphomas is a CpG-type translocation, but the t(11;18)(q21;q21)/API2-MALT1 translocation in MALT lymphomas is not. Blood 115, 3640–3641, author reply 3641–3642. [PubMed: 20430965]

Tsai AG, Yoda A, Weinstock DM, and Lieber MR (2010b). t(X;14)(p22;q32)/t(Y;14)(p11;q32) CRLF2-IGH translocations from human B-lineage ALLs involve CpG-type breaks at CRLF2, but CRLF2/P2RY8 intrachromosomal deletions do not. Blood 116, 1993–1994. [PubMed: 20847213]

Ueda Y, Liao D, Yang K, Patel A, and Kelsoe G (2007). T-independent activation-induced cytidine deaminase expression, class-switch recombination, and antibody production by immature/transitional 1 B cells. J. Immunol 178, 3593–3601. [PubMed: 17339456]

Umiker BR, McDonald G, Larbi A, Medina CO, Hobeika E, Reth M, and Imanishi-Kari T (2014). Production of IgG autoantibody requires expression of activation-induced deaminase in early-developing B cells in a mouse model of SLE. Eur. J. Immunol 44, 3093–3108. [PubMed: 25044405]

Walsh C, and Xu G (2006). Cytosine methylation and DNA repair. In Basic Mechanisms, Methylation DNA, ed. (Springer), pp. 283–315.

Watts JA, Burdick J, Daigneault J, Zhu Z, Grunseich C, Bruzel A, and Cheung VG (2019). cis Elements that Mediate RNA Polymerase II Pausing Regulate Human Gene Expression. Am. J. Hum. Genet 105, 677–688. [PubMed: 31495490]

Wei L, Chahwan R, Wang S, Wang X, Pham PT, Goodman MF, Bergman A, Scharff MD, and MacCarthy T (2015). Overlapping hotspots in CDRs are critical sites for V region diversification. Proc. Natl. Acad. Sci. USA 112, E728–E737. [PubMed: 25646473]

Welzel N, Le T, Marculescu R, Mitterbauer G, Chott A, Pott C, Kneba M, Du MQ, Kusec R, Drach J, et al. (2001). Templated nucleotide addition and immunoglobulin JH-gene utilization in t(11;14) junctions: implications for the mechanism of translocation and the origin of mantle cell lymphoma. Cancer Res 61, 1629–1636. [PubMed: 11245476]

Wiemels JL, Leonard BC, Wang Y, Segal MR, Hunger SP, Smith MT, Crouse V, Ma X, Buffler PA, and Pine SR (2002). Site-specific translocation and evidence of postnatal origin of the t(1;19) E2A-PBX1 fusion in childhood acute lymphoblastic leukemia. Proc. Natl. Acad. Sci. USA 99, 15101–15106. [PubMed: 12415113]

Yu K, and Lieber MR (2019). Current insights into the mechanism of mammalian immunoglobulin class switch recombination. Crit. Rev. Biochem. Mol. Biol 54, 333–351. [PubMed: 31509023]

Yu K, Chedin F, Hsieh C-L, Wilson TE, and Lieber MR (2003). R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. Nat. Immunol 4, 442–451. [PubMed: 12679812]

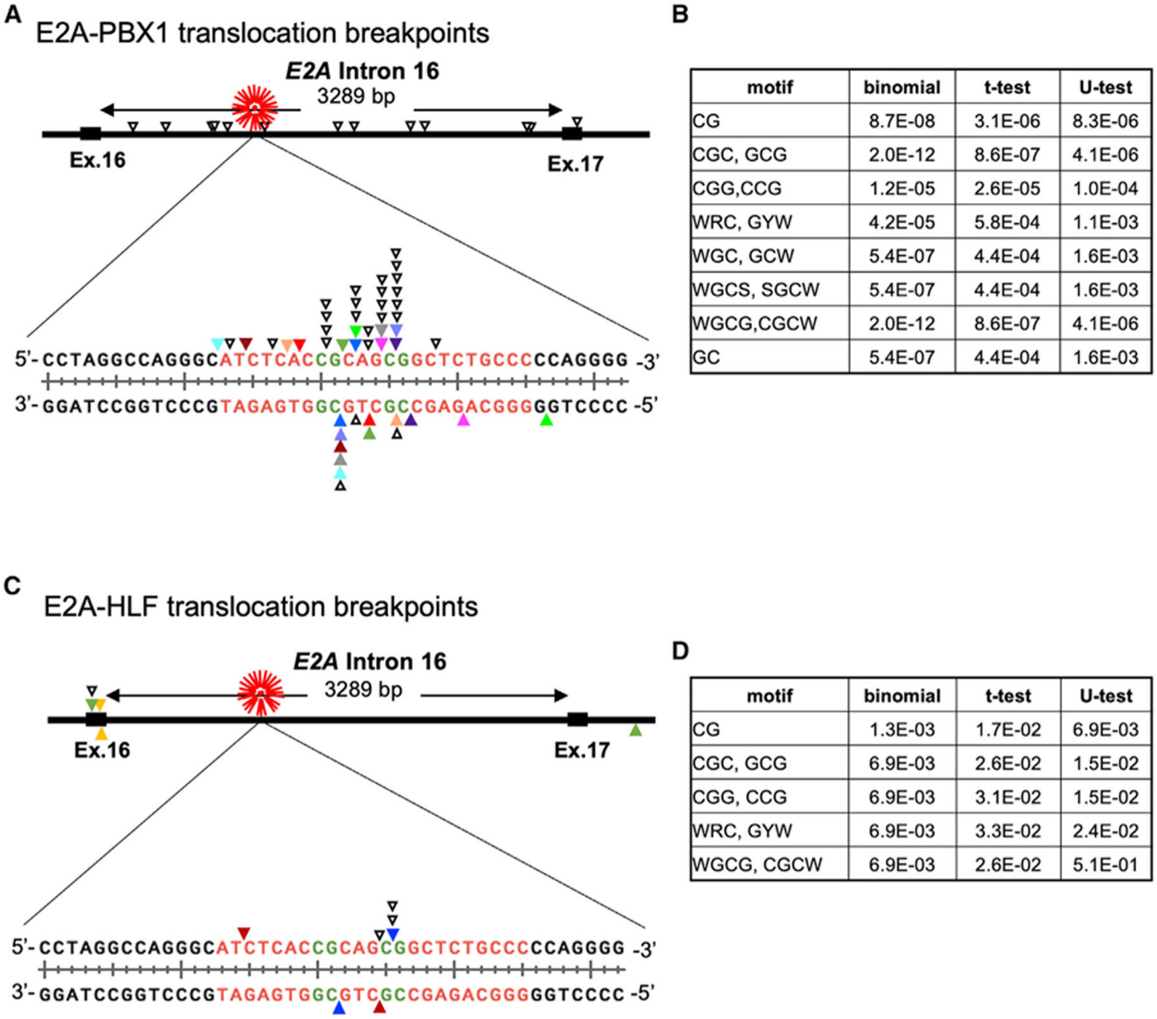
Yu K, Huang F-T, and Lieber MR (2004). DNA substrate length and surrounding sequence affect the activation-induced deaminase activity at cytidine. J. Biol. Chem 279, 6496–6500. [PubMed: 14645244]

Yu K, Roy D, Huang FT, and Lieber MR (2006). Detection and structural analysis of R-loops. Methods Enzymol 409, 316–329. [PubMed: 16793409]

Zhao B, Rothenberg E, Ramsden DA, and Lieber MR (2020). The molecular basis and disease relevance of non-homologous DNA end joining. Nat. Rev. Mol. Cell Biol 21, 765–781. [PubMed: 33077885]

Highlights

- Chromosomal translocation breaks in human B cells often occur at or near CpG sites
- Probing on live cells reveals single-stranded DNA character in the E2A fragile zone
- Purified AID targeting of the E2A fragile zone is transcriptionally stimulated
- High C-string density increases AID action in the 23 bp E2A fragile zone


A
E2A-PBX1 translocation breakpoints
E2A Intron 16
3289 bp
Ex.16
Ex.17
5'-CCTAGGCCAGGGCATCTCACCGCAGCGGCTCTGCCCCCAGGGG -3'
3'-GGATCCGGTCCCGTAGAGTGGCGTCGCCGAGACGGGGGTCCCC -5'
B
C
E2A-HLF translocation breakpoints
E2A Intron 16
3289 bp
Ex.16
Ex.17
5'-CCTAGGCCAGGGCATCTCACCGCAGCGGCTCTGCCCCCAGGGG -3'
3'-GGATCCGGTCCCGTAGAGTGGCGTCGCCGAGACGGGGGTCCCC -5'
D

motif	binomial	t-test	U-test
CG	8.7E-08	3.1E-06	8.3E-06
CGC, GCG	2.0E-12	8.6E-07	4.1E-06
CGG,CCG	1.2E-05	2.6E-05	1.0E-04
WRC, GYW	4.2E-05	5.8E-04	1.1E-03
WGC, GCW	5.4E-07	4.4E-04	1.6E-03
WGCS, SGCW	5.4E-07	4.4E-04	1.6E-03
WGCG,CGCW	2.0E-12	8.6E-07	4.1E-06
GC	5.4E-07	4.4E-04	1.6E-03

motif	binomial	t-test	U-test
CG	1.3E-03	1.7E-02	6.9E-03
CGC, GCG	6.9E-03	2.6E-02	1.5E-02
CGG, CCG	6.9E-03	3.1E-02	1.5E-02
WRC, GYW	6.9E-03	3.3E-02	2.4E-02
WGCG, CGCW	6.9E-03	2.6E-02	5.1E-01

Figure 1. The E2A breakpoints cluster to a 23 bp zone in *E2A* intron 16 in patients

(A) Distribution of E2A breakpoints in patients with E2A-PBX1 translocations. Sixty breakpoints from 49 patients were plotted as triangles. The nucleotide sequence in the 23 bp E2A fragile zone is displayed in red, with the two CpG sites in green. Triangles above and below the E2A sequence are sequenced from derivative chromosomes 19 and 1, respectively. Solid triangles with matching colors denote breakpoints on the pair of chromosomes from the same patient for a reciprocal translocation, when sequences of both junctions are available. For all remaining patients with only one of the derivative chromosomes sequenced, a hollow black triangle is used to mark the single breakpoint.

(B) Statistical analysis of sequence motifs near E2A breakpoints in patients with E2A-PBX1 translocations. Statistical analyses on the 60 E2A-PBX1 translocation breakpoints were performed to measure their proximity to more than 70 DNA motifs.

(C) Distribution of E2A breakpoints in patients with E2A-HLF translocations. Twelve breakpoints from 8 patients were plotted as triangles and illustrated in the same manner as in (A). Triangles above and below the sequence are sequenced from derivative chromosomes 19 and 17, respectively.

(D) Statistical analysis of sequence motifs near E2A breakpoints in patients with E2A-HLF translocations. Analyses were performed to measure the proximity of the 12 E2A-HLF translocation breakpoints to more than 70 DNA motifs.

For (B) and (D): W = A or T; Y = C or T; S = G or C. See STAR Methods for the details of the statistical analyses. See also Figure S1 and Tables S2 and S3.

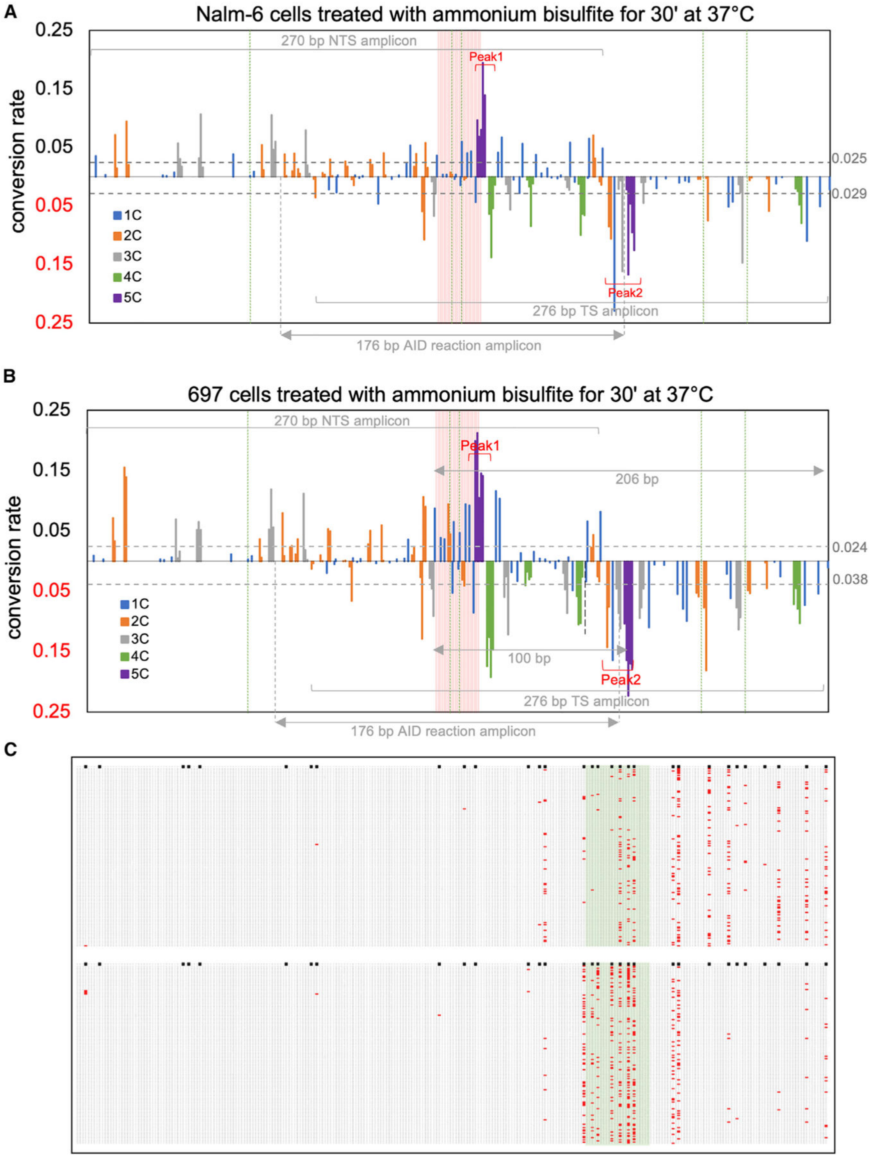


Figure 2. Whole-cell native bisulfite sequencing identifies sites with single-stranded DNA character in the 23 bp E2A fragile zone and the downstream region

(A)Native ammonium bisulfite sequencing results from live Nalm-6 pre-B cells.

(B) Native ammonium bisulfite sequencing results from live 697 pre-B cells.

(C) Exemplary E2A NTS molecules in Nalm-6 cells (top panel) and 697 cells (bottom panel) treated with native ammonium bisulfite. A collection of 160 unique molecules sequenced from the NTS of the pre-B cells treated with native bisulfite is shown in each panel. Each row represents the sequence of a unique molecule in a 5’ to 3’ direction. The single Cs on the reference sequence of the NTS are marked by black squares in the top row of each panel. The C to T conversions at single C sites in each molecule are shown as red squares. The 23 bp E2A fragile zone is highlighted in green.

For (A) and (B): the total length of the E2A region examined is 389 bp with 157 bp of overlapping region between the NTS and the TS. The x axis denotes every base position of the E2A NTS along the 389 bp region examined. The 23 bp E2A fragile zone is highlighted with a salmon background. Cytosine in the CG motif is marked with a green dashed line. The conversion rate of cytosines on the E2A NTS is shown above the x axis, and the conversion rate of cytosines on the E2A TS is shown below the x axis. The conversion rate of cytosines in different lengths of C-strings is represented in different colors as indicated in the figure. The average native bisulfite conversion rate of the single Cs located outside of the 23 bp E2A fragile zone is marked with gray dashed lines with numbers labeled on the right to serve as a reference point for the baseline conversion rate. Two major bisulfite peaks, peak 1 (at the edge of the 23 bp E2A fragile zone) and peak 2 (located 100 bp downstream), are indicated by red brackets. The location of the 176 bp region examined in the AID deamination activity experiments is indicated by the gray line with double arrows at the bottom of the figure for easy reference to the data presented in Figure 4. See also Figure S2 and Table S4.

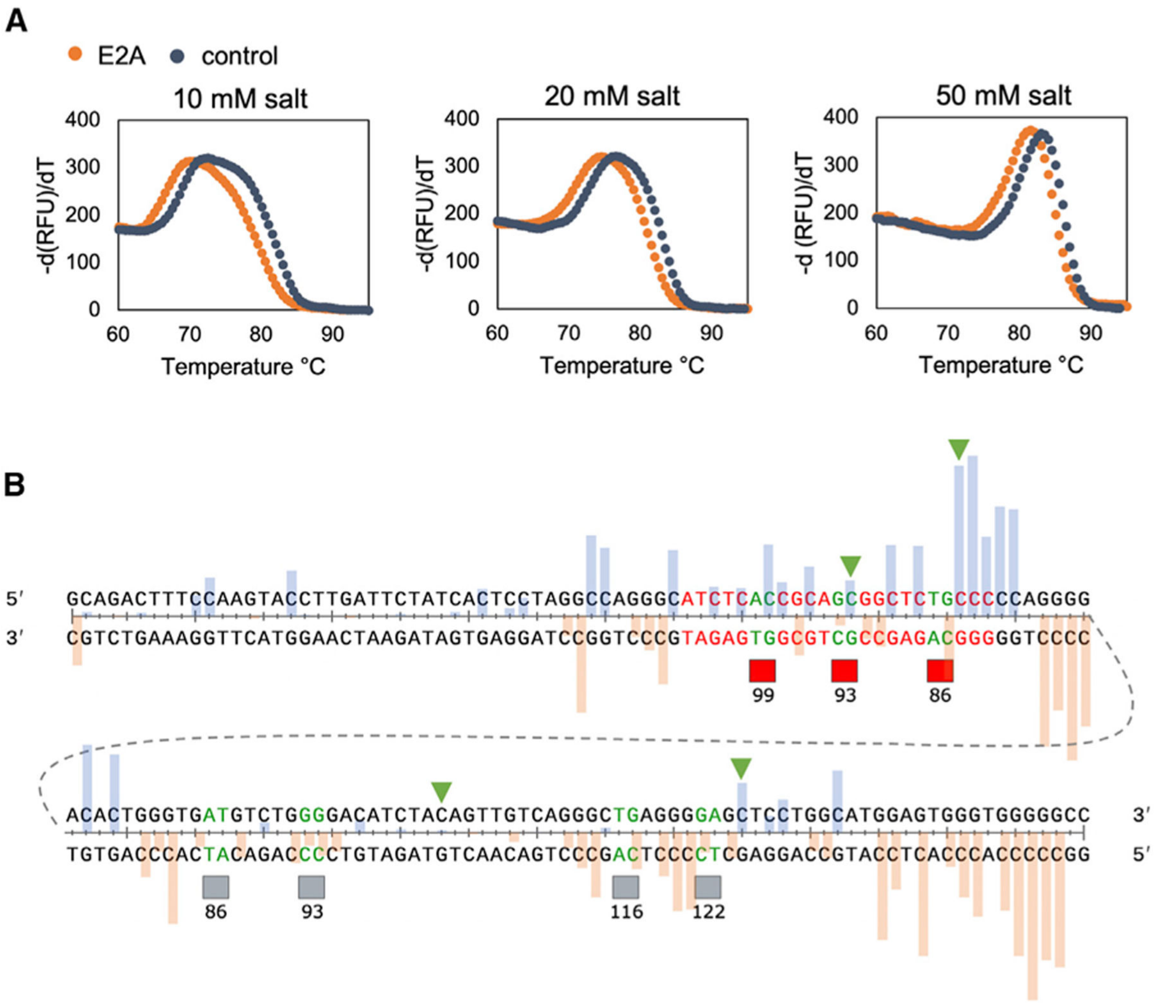


Figure 3. Melting curve analysis and nuclease P1 assay indicate thermal fluctuation within the 23 bp E2A fragile zone

(A) Melting curves show decreased DNA stability of the E2A fragile zone. Melting curve analyses of the 27 bp E2A and control substrates were performed in buffers containing 10, 20, and 50 mM sodium phosphate as indicated. The orange and gray dots represent the melting curves of the E2A duplex and the control duplex, respectively.

(B) Illustration of P1 cutting sites in the 23 bp E2A fragile zone and the surrounding regions. The nucleotide sequence of the 150 bp P1 substrate including the 23 bp E2A fragile zone (shown in red) and the surrounding regions is illustrated with the nucleotides at which P1 cuts in green. Under the nucleotide sequence, P1 cut sites in the 23 bp E2A fragile zone are indicated by red boxes, with the size of the DNA fragment below each box, as shown in Figure S3. Similarly, P1 cut sites downstream of the 23 bp E2A fragile zone are indicated by gray boxes, with the fragment size indicated below each box. To provide ease of integration with findings of other assays, the following results are also illustrated: the AID mutation sites observed on the E2A substrate without transcription and without RNase A treatment (Figure 4B) are marked with green arrow-heads. The native bisulfite reactivity from live 697 pre-B cells (Figure 2B) within the 150 bp P1 substrate is shown with light blue bars for the NTS and light orange bars for the TS.

See also Figure S3 and Table S5.

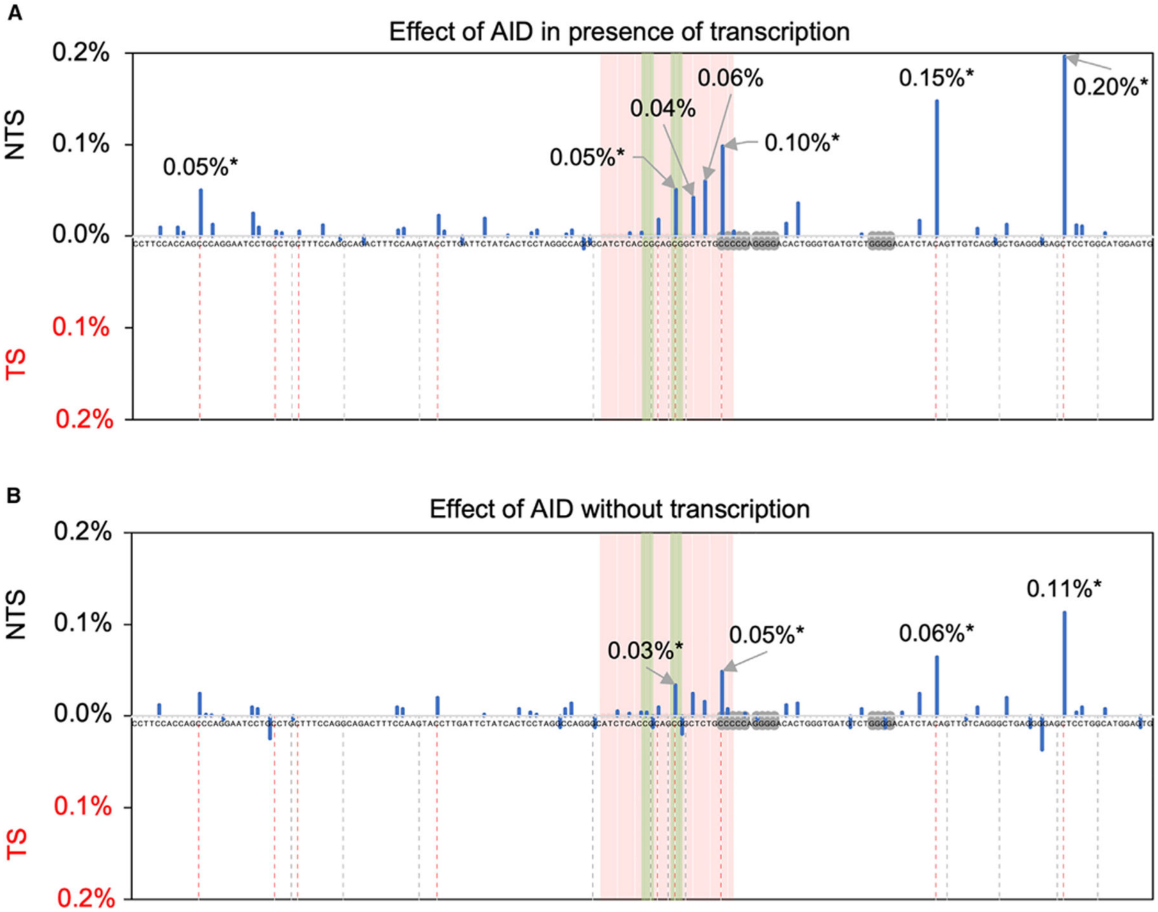


Figure 4. AID deaminates cytosines on the NTS within the 23 bp E2A fragile zone and the downstream region with and without transcription

(A) AID deamination activity in the presence of transcription (without RNase A treatment). The C to T mutation rates with appropriate background (AID-free reaction with transcription and without RNase A treatment) subtracted are presented in the histogram.

(B) AID deamination activity without transcription (without RNase A treatment). The C to T mutation rates in AID-free reaction without transcription and without RNase A treatment are used for background subtraction.

The 176 bp E2A NTS sequence in a 5’ to 3’ direction is shown in the middle of the graph, with C to T mutation rates of NTS shown above the sequence and that of TS shown below the sequence, after appropriate background is subtracted. The scale of the mutation rate is as indicated on the left of the graph as a percentage. The mutation rates at high-AID deamination sites are as annotated, and AID hotspot sites are marked with an asterisk. The 23 bp E2A fragile zone is broadly colored in salmon. AID hotspot sites (WRC and CGC; Yu et al., 2004) on NTS and TS are aligned with red and gray dashed lines, respectively. The CpG sites are highlighted by the green zones. The sequence motifs covered by gray boxes on the x axis are the ones disrupted in mut1 and mut2, as in Figure 6, for easy reference. See also Figure S4 and Table S6.

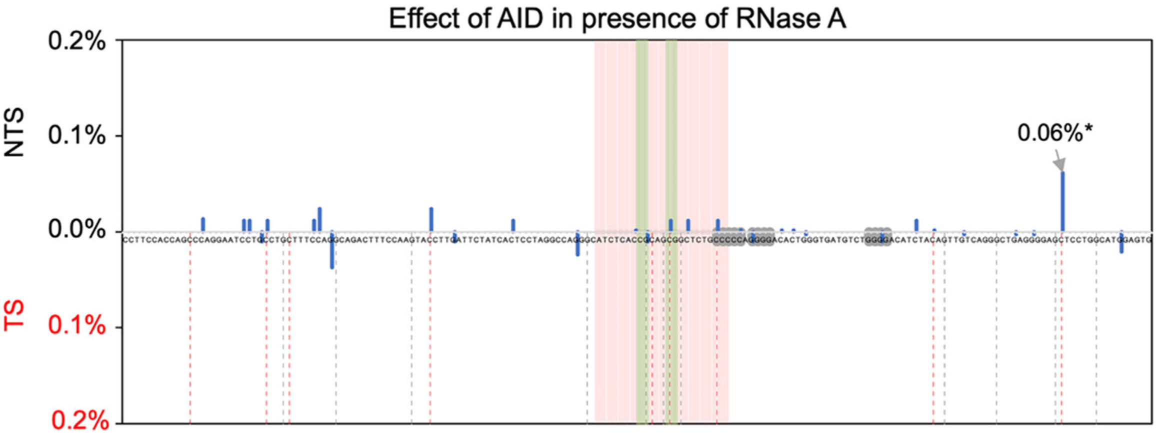


Figure 5. Presence of RNase A during transcription results in markedly decreased AID deamination activity in the E2A fragile zone

The C to T mutation rates reflect AID deamination activity in the presence of RNase A during transcription on the 176 bp E2A substrate with the background (AID-free reaction with transcription and with RNase A) subtracted. The figure is shown in the same configuration as described in Figure 4. See also Figure S5 and Table S6.

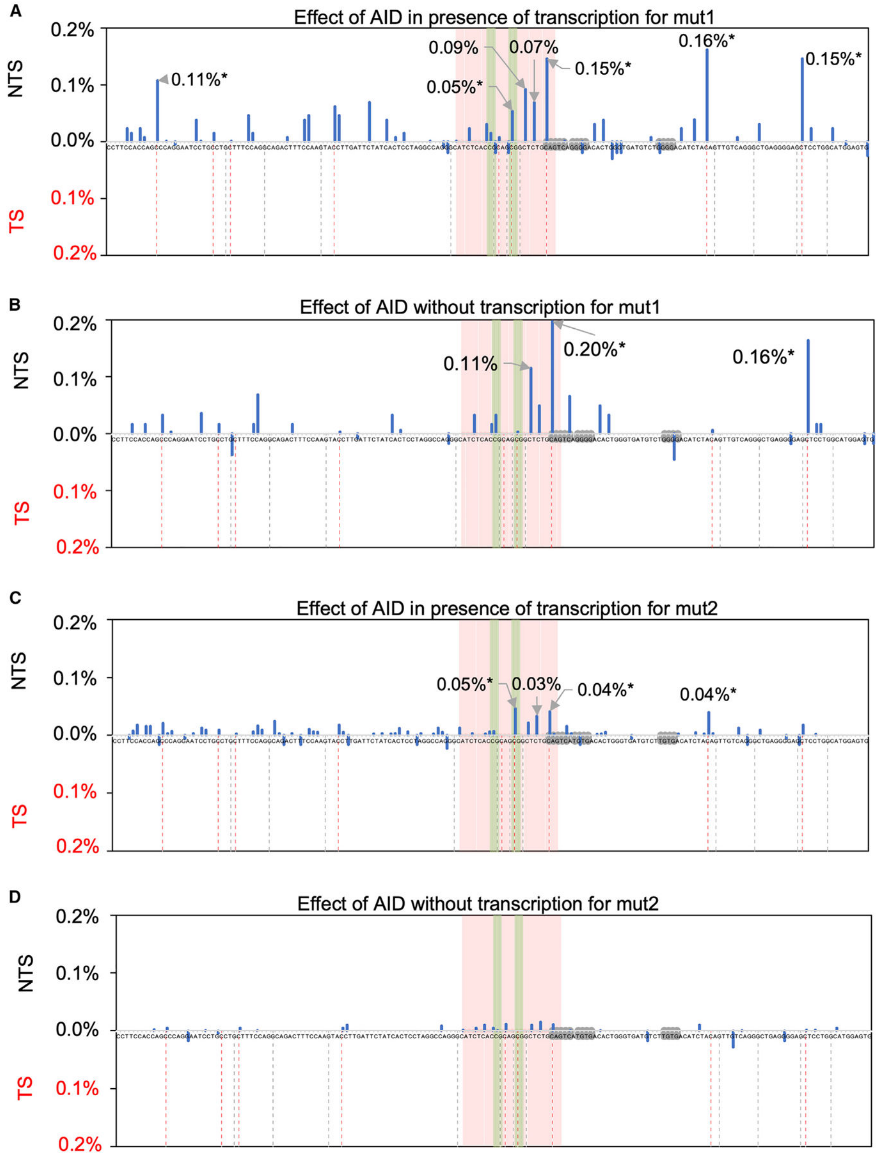


Figure 6. AID has decreased deamination activity on the NTS of the mut2 substrate, but not the mut1 substrate, with and without transcription

(A) AID deamination activity on mut1 substrate in the presence of transcription (no RNase A). The C to T mutation rates presented are with the background (AID-free reaction with transcription and without RNase A) subtracted.

(B) AID deamination activity on mut1 substrate in the absence of transcription (no RNase A). The C to T mutation rates presented are with the background (AID-free reaction without transcription and without RNase A) subtracted.

(C) AID deamination activity on mut2 substrate in the presence of transcription (no RNase A). The C to T mutation rates presented are with the background (AID-free reactions with transcription and without RNase A) subtracted.

(D) AID deamination activity on mut2 substrate in the absence of transcription (no RNase A). The C to T mutation rates presented are with the background (AID-free reaction without transcription and without RNase A) subtracted.

Other than the NTS sequences of mut1 (A and B) or mut2 (C and D) in a 5’ to 3’ direction shown in the middle of the graph, the remaining of the figure is shown in the same configuration as described in Figure 4. See also Figure S6 and Table S6.

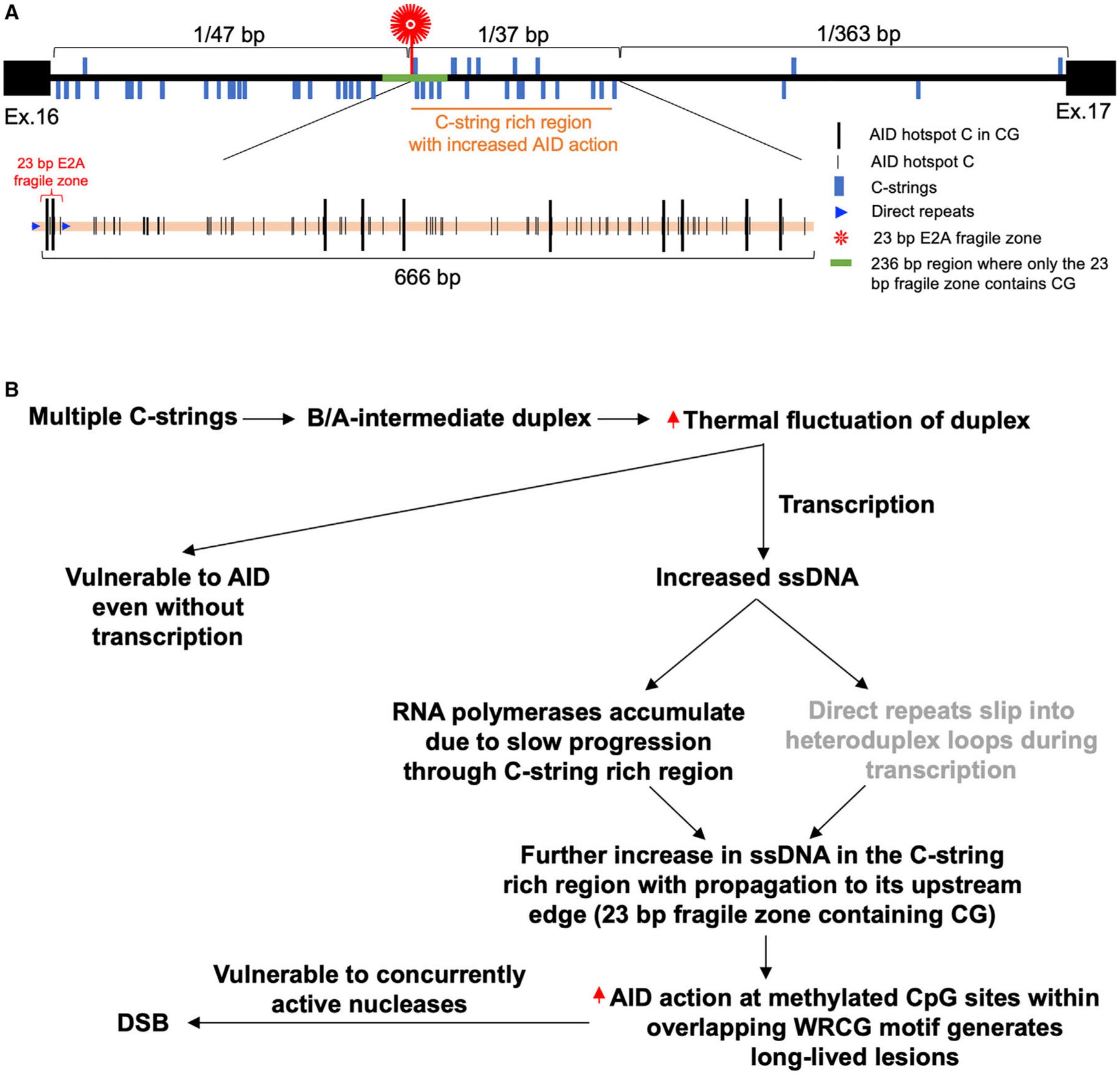

A
1/47 bp
1/37 bp
1/363 bp
Ex.16
Ex.17
C-string rich region
with increased AID action
23 bp E2A
fragile zone
666 bp
AID hotspot C in CG
AID hotspot C
C-strings
Direct repeats
23 bp E2A fragile zone
236 bp region where only the 23 bp fragile zone contains CG
B
Multiple C-strings
B/A-intermediate duplex
Thermal fluctuation of duplex
Transcription
Vulnerable to AID even without transcription
Increased ssDNA
RNA polymerases accumulate due to slow progression through C-string rich region
Direct repeats slip into heteroduplex loops during transcription
Further increase in ssDNA in the C-string rich region with propagation to its upstream edge (23 bp fragile zone containing CG)
AID action at methylated CpG sites within overlapping WRCG motif generates long-lived lesions
Vulnerable to concurrently active nucleases
DSB

Figure 7. Factors and order of steps important for breakage at the E2A fragile zone

(A) Factors contributing to the clustered E2A breakage. The *E2A* intron 16 is shown as a black line between exon 16 and exon 17. The E2A fragile zone is marked by the red asterisks above the black line. The green portion of the top horizontal line is a 236 bp portion devoid of CpG sites except for the two within the 23 bp fragile zone. C-strings with a length of 4 and longer are shown as vertical blue lines and annotated above the black line for the NTS and below the black line for the TS. The density of C-strings in each of the three regions in *E2A* intron 16 is shown above the brackets. The enlarged view of the 666 bp region with high C-string density on both strands is illustrated below intron 16 by an orange horizontal line. Cytosines in AID hotspot motifs in this region are shown as thin vertical black lines. Cytosines in both AID hotspot motifs and CpG sites (therefore WRCG) are shown in bold vertical black lines. The two blue arrowheads on the sequence represent the direct DNA repeats flanking the E2A fragile zone.

(B)Steps leading to E2A breakage. DNA regions with high C-string density are in a B/A-intermediate DNA conformation, which has increased DNA thermal fluctuation, and are vulnerable to AID deamination. Transcription through the C-string-rich region increases its single-stranded character by the accumulation of RNA polymerases and by the slippage between DNA direct repeats. The cytosines in the WRCG motif within this region are preferred targets of AID deamination, which leads to persistent DNA lesions when cytosines are methylated. The long-lived DNA lesions at methylated CpG sites within AID hotspots are subject to nuclease activities (RAG or activated Artemis), resulting in DSBs.