SUPPLEMENTARY TEXT

Timing of B cell translocations

Pro-B and pre-B stage

The translocations in early B cells typically have random junctional additions that reveal what stage of B cell development that the translocation occurred in. The designation for junctional additions is N-nts, which are C/G-rich additions (often with runs of the same nt) by terminal deoxynucleotidyl transferase (TdT) during V(D)J recombination in both early B and early T cells. In most of the lymphoid translocations, the N-nt feature is seen at the junctional sequences of the breaks that occur at the non-IGH loci, such as *E2A*, *MALT1*, and *CRLF2*. Among the 72 junctional sequences in E2A involved translocations that we have analyzed, 66 of them contain N-nts, and the remaining ones show no nucleotide additions. About 60% of the inserted nucleotides (458 out of 773) are Cs and Gs, and 46 out of 72 (65%) junctional sequences contain at least three consecutive Cs or Gs. N-nts are observed in all eight patients with IGH-MALT1 translocation with four of them containing consecutive Cs or Gs. Eighteen out of 19 junctional sequences around CRLF2 breakpoints in CRLF2-IGH translocations also contain N-nts, consistent with TdT activity.

A second junctional addition type, called T-nts, are a copy of 3 to 15 nt from either of the two DNA ends involved in the translocation. Both N-nts and T-nts are observed in fragile zones of *BCL2* and *BCL1*, and these are usually but not always distinguishable. Over 97% of the BCL1 breakpoints contain inserted junctional sequences, and 57% of the 1358 inserted nucleotides are Cs or Gs with frequent presence of C/G-strings, typical of TdT addition. T-nts of 8 to 12 nts are observed in 9

out of the 104 BCL1 MTC breakpoints, exhibiting mismatches with the germ-line sequences from MTC or IGH regions (Welzel et al. 2001). Ten out of 38 breakpoints located outside of BCL1 MTC region show T-nts of ≥ 8 nts. Over 95% of the junctional sequences of *BCL2* show nucleotide additions, the majority of which are random nucleotides (not templated), a characteristic of TdT activity. One study reported T-nt insertions at the BCL2-IGH junctions in 30% of the follicular lymphoma (FL) patients (Jager et al. 2000). The two different nucleotide addition patterns in *BCL1* and *BCL2* may indicate the involvement of two different breakage/repair mechanisms. The addition of T-nts may indicate that pol μ and pol λ had a longer time window to modify the DNA ends. It is possible that an alternative end joining (aEJ) pathway may be responsible for some T-nts (Carvajal-Garcia et al. 2020). We favor the view that all the N-nts and a majority of the T-nts are added during NHEJ for the following reasons: (a) most of the junctions have TdT additions, indicating that TdT is present and indicating a typical NHEJ event in lymphoid cells (Gauss and Lieber 1996); (b) pol μ and λ can generate direct repeats (DR) and inverted repeats (IR), which are the essential feature of T-nts (Maga and Hübscher 2003); (c) aEJ typically generates at least 2 nts of microhomology at both ends of the junctional addition, and this is rarely observed in lymphoid cells (Carvajal-Garcia et al. 2020).

The presence of N-nts in the junctional regions of *E2A*, *BCL1*, *BCL2*, and *MALT1* indicates that these chromosomal translocations arise during the pro-B or pre-B cell stage. For *BCL1*, *BCL2* and *MALT1*, this is consistent with the partner DSB arising at the *IGH* locus during V(D)J recombination (which is a pro-B/pre-B cell event).

Mature B stage

Translocations involving *IGH* and *BCL6* or *MYC* occur in much larger zones and do not show features of TdT addition. Among 63 cases with *BCL6* translocated to

immunoglobulin loci with sequence information, 22 contain 1-5 nt microhomology, 23 contain random insertions, and 18 have no identifiable insertions or microhomology. Among the 58 cases with BCL6 translocated to non-immunoglobulin loci with sequences available, 29 have 1 to 4 nts microhomology, 9 contain random insertions, and 20 have no insertions or microhomology. The low percentage of N-nts presence in the junctional sequences of *BCL6* translocations is consistent with the view that these translocations are more likely to occur in mature B cells rather than in pro-B cells. Among the 177 *MYC* breakpoints, 33 contain nucleotide insertions of 1 nt to 33 nt in length with most of them less than 5 nt; 92 have microhomology of 1 to 5 nts from either *MYC* or *IGH* sequences with no insertions; and the remaining show no insertions or no microhomology (when the *IGH* break junctional sequence is available for inspection).

The characteristic of *BCL6* and *MYC* junctional sequences is very different from that of *BCL1*, *BCL2*, *MALT1*, and *E2A*, which mostly contain N-nts. The mutation patterns of BCL6 and MYC (caused by SHM) is consistent with the distribution of breakpoints, indicating AID activity in geminal center B cells is the cause of the breaks (Lu, Pannunzio, et al. 2015).

Biological aspects of major translocation events in human pro-B and pre-B cells

BCL2-IGH translocations

The BCL2-IGH translocation (Fig. 2A) is found in 50% of non-Hodgkin's lymphomas (NHL) including over 80% FL and ~20% of diffuse large B cell lymphomas (DLBCL) (Finnon et al. 1995; Buchonnet et al. 2000). It results from illegitimate rearrangement

between the *BCL2* gene and D_H/J_H gene segments during V(D)J recombination in early B cells (Fig. S4A). No new chimeric protein is generated from the translocation event because the BCL2 breakpoints are mainly located at the 3' UTR and its downstream intergenic region. The t(14;18) event generates a hybrid transcript that consists of a major portion of the 5' moiety of the BCL2 mRNA fused to the enhancer region of the immunoglobulin gene and increases the expression level of BCL2 (Cleary et al. 1986). The BCL2 protein, known as a potent apoptosis inhibitor, blocks programmed cell death and promotes cell survival (Reed 1994). The increased BCL2 level in ALL patients is associated with poor response to chemotherapy (Campos et al. 1993).

BCL1-IGH translocations (also called CCND1-IGH)

The t(11;18) BCL1-IGH translocation (Fig. 2B) mainly occurs in mantle cell lymphomas, but also are found in plasma cell leukemias (PCL), intermediate lymphocytic lymphomas (ILL), and chronic lymphocytic leukemias (CLL) (Raffeld and Jaffe 1991; Rimokh et al. 1993; Resnitzky et al. 1996; Shimazaki et al. 1997; Remstein et al. 2000). The BCL1 breakpoints are scattered within a 344 kb region between *CCND1* and *MYEOV* gene. The translocation event does not lead to new chimeric proteins but may exchange enhancers between the *IGH* locus and the *BCL1* gene (Fig. S4B). The *IGH* intronic enhancer is translocated to the upstream region of *CCND1* gene, stimulating the expression of CCND1. The dysregulation of CCND1 due to the translocation event forces the cells to enter the S phase of the cell cycle and play a key role in the pathogenesis of B cell lymphomas.

E2A-PBX1 and E2A-HLF translocations

The E2A-PBX1 translocation (Fig. 2C) occurs in 5% of paediatric ALL patients (Foa et al. 2003). E2A is an important regulator of lymphocyte differentiation and maturation

with high expression in the lymphoid system (Bain et al. 1994; Kee et al. 2000). The E2A protein is composed of two activation domains and a basic helix-loop-helix domain. PBX1, usually not expressed in lymphoid cells, encodes a HOX family transcription factor which is made up of a C-terminal homeodomain and a dimerization domain. The PBX1 C-terminal homeodomain usually interacts with other HOX family transcription factors to regulate gene expression (Knoepfler and Kamps 1995). The derivative E2A-PBX1 gene after chromosomal translocation containing *E2A* exon 1-16 and *PBX1* exon 3-9 leads to the expression of a novel chimeric protein that is composed of the DNA binding domain of PBX1 and the transactivation domains of E2A (Fig. S4C) (Nourse et al. 1990). The chimera can continuously activate the expression of PBX1 targeted genes in the B lymphoid compartment, where PBX1 is not expressed under physiological conditions (LeBrun and Cleary 1994). *In vitro* and *in vivo* studies both have shown that this oncogenic translocation is capable of cellular transformation and tumorigenesis (Kamps et al. 1991; Lin et al. 2019; Pi et al. 2020). The other derivative gene (composed of 5' of *PBX1* and 3' of *E2A*) is usually either intact and then silent in B cells or lost from the leukemia cells.

IGH-MALT1 and MALT1-API2 translocations

MALT1 has two common translocation partners, *IGH* and *API2*. The t(14;18) IGH-MALT1 translocation (Fig. 2D) occurs in 10% of mucosa-associated lymphoid tissue (MALT) lymphomas (Murga Penas et al. 2010), which takes up 8% of all NHL (Troppan et al. 2015). The breakpoints of *MALT1* are located upstream of the coding exons in the intergenic region. The tumor cells with IGH-MALT1 translocation showed dysregulated expression of MALT1 and the downstream *BCL10* gene (Ye et al. 2005).

The t(11;18) API2-MALT1 translocation occurs in 50% of MALT lymphomas (Lucas et al. 2001). The breakpoints of *MALT1* in API2-MALT1 translocation are

widely scattered within a 29 kb region in *MALT1*. The translocation event leads to a fusion protein that fuses the N-terminus of *API2* gene to the C-terminus of *MALT1* gene (Morgan et al. 1999). API2 is one of the inhibitors of apoptosis proteins. The resulted API2-MALT1 chimera was reported to self-oligomerize via the N terminus domain from API2 and is able to activate NF-kB pathway (Lucas et al. 2001).

IGH-CRLF2 and P2RY8-CRLF2 translocations

CRLF2 rearrangement is seen in 14% children with B-cell precursor acute lymphoblastic leukemias (B-ALL) and in 63% of paediatric ALL patients with Down Syndrome (Harvey et al. 2010; Hertzberg et al. 2010). Rearrangement of *CRLF2* is seen in 50% of the Ph-like acute lymphoblastic leukemia (Ph-like ALL), a high-risk group of B-cell ALL that lacks BCR-ABL1 fusion (Herold et al. 2017; Jain et al. 2017). An intrachromosomal deletion between of the pseudoautosomal region 1 of chromosomal X/Y occurs in 7% of B cell ALL patients, resulting in the juxtaposition of the first noncoding exon of *P2RY8* with the coding region of *CRLF2* (Mullighan et al. 2009). P2RY8-CRLF2 translocation leads to the expression a new fusion protein and is associated with JAK kinase mutations, both of which contribute to the leukemogenesis of B-progenitor ALL (Mullighan et al. 2009; Russell et al. 2009; Yoda et al. 2010). The breakpoints in this interstitial deletion event are located near heptamer sequences of the RAG complex motif upstream of *CRLF2* and within intron 1 of *P2RY8*, indicating the erroneous V(D)J recombination as the cause of the translocation (Tsai et al. 2010). The t(X;14) IGH-CRLF2 (Fig. 2E) occurs in 9% of paediatric pre-B ALL patients (Harvey et al. 2010). The breakpoints of *CRLF2* in this translocation event are scattered upstream of *CRLF2* gene in the intergenic region with a small cluster, resulting in enhanced expression of CRLF2 gene after fusion with the *IGH* locus. Like P2RY8-CRLF2 event, patients with CRLF2 overexpression also have high JAK kinase mutation

rate, which contribute to the extremely poor treatment outcomes and increased B cell leukemogenesis (Harvey et al. 2010).

Biological aspects of major translocation events in human mature B cells

IGH-BCL6 translocations

IGH-BCL6 translocations (Fig. 2F) occur in around ~10% of NHL (Bastard et al. 1992; Deweindt et al. 1993). Besides the *IGH* locus, *BCL6* has been reported to translocate to many other genes including *PIM1*, *RHOH*, *HSPCA*, and *TFRC* etc. in B cell tumors (Akasaka et al. 2000; Chen et al. 2006). Overall, BCL6 translocations in B cell NHL are observed in 5-15% in FL, 20-40% in DLBCL, and 20% acquired immunodeficiency syndrome (AIDS)-associated DLBCL (Ohno 2011).

BCL6 gene encodes a sequence-specific transcription factor of Krüppel-like subfamily. BCL6 protein functions to repress transcription from promoters containing its DNA-binding site and is exclusively expressed in geminal center B cells (Ohno 2004, 2011; Basso and Dalla-Favera 2012). It is key for the development of germinal center B cells and follicular helper T cells (Kitano et al. 2011). The breakpoints of *BCL6* span over its promoter, non-coding exon 1, and intron 1. The rearrangement of *BCL6* with the *IGH* locus does not generate new fusion proteins but juxtaposes the *IGH* upstream sequence to *BCL6* in the same transcriptional orientation (Fig. S4D). Similar, for BCL6 translocations involving other genomic loci, the coding region of *BCL6* is usually fused to the promoter sequence of the partner genes, which leads to the aberrant expression of BCL6. These translocation events with deregulated BCL6 expression usually repress terminal B cell differentiation, affect antibody response, and then finally contribute to malignant lymphomas (Offit et al. 1994; Wagner et al. 2011).

IGH-MYC translocations

Around 80%-90% Burkitt's lymphoma (BL) cases contain t(8;14) IGH-MYC translocation (Fig. 2G), making it a hallmark of BL (Hecht and Aster 2000; Boxer and Dang 2001). *MYC* gene contains three exons. Its exon 1 contains two promoters and is noncoding (Battey et al. 1983). The breakpoints of *MYC* are scattered from the region upstream of *MYC* to its intron 1. The translocation event makes the two genes fused in a head-to-head manner in which the *MYC* exon 2 and 3 are joined to 5' of *IGH* region (Fig. S4E). The normal *MYC* allele is usually silent in BL, and MYC protein is only expressed from the translocated *MYC* allele on derivative chromosome 14 (Hayday et al. 1984; Cory 1986). The expression of the translocated MYC is driven by the upstream P1 promoter instead of P2 promoter, leading to enhanced MYC expression that drives the cell proliferation in BL (Strobl et al. 1993). The breakpoints on *IGH* gene occur mainly in class switch regions, with a few located in V(D)J recombination region or sequence compatible with V(D)J recombination, or other regions/unknown. The cleavage sites of *IGH* indicate the involvement of either RAG complex in V(D)J recombination or AID (switch regions) in CSR.

Motif analyses of the breakpoints at B cell oncogenes

BCL2 breakpoints at CG and AID motifs

A total of 551 BCL2 breakpoint sequences in BCL2-IGH translocation are currently in the literature that we have compiled in our database. The breakpoints of the *IGH* loci are usually located at the coding ends of the D_H and J_H segments (Fig. S4A), which are from failed DNA ends joining after the RAG complex resulted DNA breakage in V(D)J recombination. In contrast, the statistical analysis at the BCL2 breakpoints does not

show significant proximity to CAC motif (Tsai et al. 2008). This indicates that other mechanisms irrelevant to the RAG complex are involved in the BCL2 breakage.

All the BCL2 breakpoints spread over an 18 kb length with three main clusters, each one being 105 to 570 bp in size. The 175 bp major breakpoint region (MBR) located at the 3' UTR of *BCL2* gene contains 487 (88%) BCL2 breakpoints (Fig. 3A). The three major peaks within the MBR of *BCL2* are all centred at a CG motif. A total of 208 (43%) BCL2 MBR breakpoints are directly (zero nts) at CG ($p = 1.8x10^{-96}$ in binomial test in Table 2). The average distance in the MBR to the CG motif is 4.4 bp, in contrast with the 11.2 bp if the breakpoints were randomly distributed ($p = 1.2 \times 10^{-42}$ in U-test). The breakpoints in MBR are also in significant proximity to the AID CGC motif ($p = 1.4 \times 10^{-48}$ in U-test), which contains the CG in motif just mentioned in nearly all cases.

Eleven (2%) BCL2 breaks occur in the 105 bp intermediate cluster region (icr) that is 19 kb downstream of the MBR (Fig. 3A). In all icr breakpoints, 73% (8) of them are directly at (zero nts away) CG ($p = 1.5x10^{-8}$ in binomial test in Table 2). The average distance of icr breakpoints to CG motif is 0.55 bp, compared with 4.7 bp if they occurred in a random distribution ($p = 8.2 \times 10^{-8}$ in U-test). The BCL2 breakpoints in icr do not show significant proximity to AID hotspot motifs.

Nineteen (3%) of BCL2 breakpoints are mapped to a 561 bp minor cluster region (mcr) that is 29 kb downstream of MBR (Fig. 3A). Fourteen (74%) of the 19 mcr breaks are directly at (zero nts away) from CG motif ($p = 5.4x10^{-18}$ in binomial test). The average distance of mcr breaks to CG is 0.6 bp versus 40 bp if they were in a random pattern ($p = 7.6 \times 10^{-13}$ in U-test in Table 2). Besides CG, the icr breakpoints are also statistically significantly close to AID CGC ($p = 4.8 \times 10^{-10}$ in U-test) and WRC ($p = 4.6 \times 10^{-6}$ in U-test) hotspot motifs, which often contain a CG (for CGC) or have a

CG at the 3' edge (for WRC).

Among 27 breaks analysed that were scattered between the BCL2 MBR, icr, or mcr, but were not within any of these three regions, 5 (19%) of them are directly at (zero nts) CG ($p = 2.2x10^{-3}$ in the binomial test).

Therefore, the statistical analyses regarding BCL2 breakpoints in MBR, icr, mcr, and the interspace regions indicate CG motif and AID are important in BCL2 breakage phase.

BCL1 breakpoints at CG and AID motifs

A total of 162 DNA sequences from BCL1 breakpoints in BCL1-IGH translocations are in the Lieber lab database (Fig. 3B). The 150 bp BCL1 major translocation cluster (MTC) located 109 kb upstream of *BCL1* gene contains 104 (64%) BCL1 breaks, with the remaining 58 breakpoints broadly scattered outside of the MTC in a 329 kb region in the 344 kb intergenic zone between *CCND1* and *MYEOV* (non-MTC breaks).

Among the 104 BCL1 breakpoints in MTC, 38 (37%) of them are directly at (zero nts) CG ($p = 7.0 \times 10^{-9}$ in the binomial test in Table 2). Over 90% (96 in 106) of the BCL1 breaks in MTC are within 8 bp to CG. The average distance of BCL1 breakpoints in MTC to CG motif is 2.6 bp, in contrast with the 7.8 bp if the breakage occurs in random pattern ($p = 1.1x10^{-12}$ in U-test in Table 2). The breakpoints do not show significant proximity to AID WRC and WGCW hotspot motifs ($p>0.1$) but are significantly close to AID CGC hotspot motif ($p = 7.0 \times 10^{-9}$ in U-test), which contains the CG motif.

Twenty-nine out of the 58 (50%) non-MTC BCL1 breakpoints are less than 5 nt from CG motif, among which 18 of them are right at CG. The CG type breaks (≤ 5 nt from CG) and non-CG type breaks (> 5 nt from CG) have different distribution across the BCL1 329 kb break region (Greisman et al. 2012). Of note,11 of the 29 CG type

non-MTC BCL2 breakpoints are at AID CGC motif ($6.7x10^{-13}$ in binomial test). The statistical analysis on the 29 non-CG type breaks shows no significant proximity to either CG motif or AID hotspot motifs.

The significant proximity of BCL1 breaks in MTC and CG type breaks in non-MTC to CG motif and AID hotspot motifs indicates the critical roles of CG and AID in their breakage. The non-CG BCL1 breaks located outside of MTC may indicate a different mechanism for their breakage.

E2A breakpoints at CG and AID motifs

A total of 60 breakpoint sequences around the rejoining sites from 49 patients (and cell lines) with the E2A-PBX1 translocations have been reported (Wiemels et al. 2002; Paulsson et al. 2007; Fischer et al. 2015; Kato et al. 2017; Hein et al. 2019). Eight patients with E2A-HLF translocations have been reported with known sequences around the rejoining sites (Hunger et al. 1992; Inaba et al. 1992; Fischer et al. 2015).

E2A breakpoints from 48 of the 49 patients (98%) with E2A-PBX1 translocations are located within the 3.3 kb *E2A* intron 16 (Fig. 3C). Surprisingly, 36 out of the 48 patients (75%) with E2A-PBX1 translocations have E2A breaks localized to a 23 bp zone in *E2A* intron 16, making the 23 bp zone > 400-fold more fragile compared with other regions within the same intron. Within the 23 bp E2A fragile zone, there are two CpG sites. Among the 47 breaks located inside the fragile zone from patients with E2A-PBX1 translocation, 39 (85%) of them are within 1 bp distance to CG motif, 28 (60%) of them being right at the CG and 11 breakpoints being 1 bp away from CpG sites. Among the 49 cases with E2A-PBX1 translocations, 11 of them have the breakpoint sequences from both derivative chromosomes reported (reciprocal translcoations). After the initial DNA breakage, the DNA ends are predominantly rejoined through NHEJ. The rejoining process always involves DNA end resection and

nucleotide insertion prior to NHEJ (Chang et al. 2017). Therefore, the breakpoints sequenced from two derivative chromosomes can provide us with a sequence window where the initial DNA breakage might have originated (Lieber 2016). The 11 patients with reciprocal E2A-PBX1 translocations provide us with an average of 5.5 bp initial breakage window ranging from 1 bp to 14 bp. All the 11 breakage windows from the reciprocal translocations include the CG motif. Statistical analyses indicate E2A breaks from E2A-PBX1 translocation are statistically significantly in proximity to the CG motif ($p = 8.3 \times 10^{-6}$ in U-test), AID WRC motif (W=A/T, R=A/G, $p = 1.0 \times 10^{-3}$ in U-test), and AID CGC motif ($p = 4.1 \times 10^{-6}$ in U-test).

Five of the 8 patients with E2A-HLF translocation have E2A breakpoints in intron 16, including two patients with reciprocal translocations. The seven breakpoints from the five patients, including two pairs of double-breakpoints and 3 single-breakpoints are all within the 23 bp E2A fragile region (Fig. 3C), which further confirms the fragility of the 23 bp zone. Six out of 7 E2A breaks (86%) within the 23 bp region in patients with E2A-HLF translocations are also directly at the CG motif ($p = 1.3 \times 10^{-3}$ in the binomial test). The E2A breakpoints in E2A-HLF translocations are also in significant proximity to CG motif ($p = 6.9 \times 10^{-6}$ in U-test) and AID WRC ($p = 2.4 \times 10^{-2}$ in U-test) and CGC ($p = 1.5 \times 10^{-2}$ in U-test) hotspot motifs.

The breakpoint sequences of *E2A* strongly suggest the critical role of CG motif and AID in the breakage phase of the E2A fragile zone.

MALT1 breakpoints at CG and AID motifs

Eight MALT1 breakpoints from IGH-MALT1 translocation are available, all of which are highly focused in an 86 bp zone that is located 1.4 kb upstream of the *MALT1* gene in the intergenic region (Fig. 3D). Seven out of the 8 MALT1 breakpoints are within 8 bp to the CG motif, 4 of which are right at CG. Statistical analyses show the

breakpoints are significantly close to CG motif ($p = 6.2 \times 10^{-3}$ in U-test in Table 2) and AID hotspot motif (WGCW, $p = 1.6 \times 10^{-3}$ in U-test).

Twenty MALT1 breakpoints in API2-MALT1 translocation are broadly located in a 29 kb region across the whole *MALT1* gene, none of which are at CpG sites ($p = 1$ in binomial test). The API2 breakpoints are scattered in a 4.5 kb region within its intron 7. The MALT1 breakpoints in API2-MALT1 translocation do not show significant proximity to AID WRC ($p = 0.5$ in U-test) and WGCW ($p= 0.8$ in U-test) hotspot motifs, but they show some propensity to AID CGC hotspot motif ($p = 3.5 \times 10^{-2}$ in U-test) than by random chance.

The junctional sequences of MALT1 breakpoints and the statistical analyses results suggest the different mechanisms behind the breakage of *MALT1* in IGH-MALT1 and in API2-MALT1 translocations. Like other fragile zones, CG and AID are critical for MALT1 breakage in IGH-MALT1 translocation but other factors may contribute to its breakage in API2-MALT1 translocation.

BCL6 and MYC breakpoints are at WRC and WGCW and a subset are at CG motif

IGH-BCL6 and IGH-MYC are two common chromosomal translocations occurring in B cell malignancies. R-loops that are kilobases in length have been described at both loci and account for the large regions in which AID-type breaks are generated at these two loci in mature B cells (Ruiz et al. 2011; Lu et al. 2013; Yang et al. 2014). Unlike the translocations that occur in early B cells, the BCL6 and MYC breakpoints are widely scattered in large zones up to 4.1 kb. Different from the fragile regions mentioned above, the 4.1 kb MYC break region and 2156 bp BCL6 break region are very CG-rich, averaging one CG per 16.8 bp for MYC and one CG per 18.9 bp for BCL6. In contrast, the CG density in MTC, MBR, icr, and mcr is 21.4 bp, 35 bp, 50.3 bp, and 93.5 bp.

The 152 BCL6 breakpoints spread over an 18 kb bp region, starting from 6.7 kb upstream of the *BCL6* gene to its intron 2 (Fig. S5A). Eight-five of 152 (56%) BCL6 breakpoints translocate to *IG* loci and 67 (44%) of them translocate to non-*IG* loci such as *PIM1*, *CIITA*, *RHOH*, and *HNRNPC*. In all 152 BCL6 breakpoints, 134 (88%) are focused within a 2156 bp region in *BCL6* intron 1, among which 81 breakpoints are from BCL6-IG translocations and 53 breakpoints are from BCL6-non-IG translocations (Fig. S5A). Fourteen of the 81 BCL6 breakpoints in BCL6-IG translocation within the 2156 bp fragile zone are right at CG ($p = 0.35$ in binomial test). They show highly significant proximity to AID hotspot motifs (for WGCW: $p = 4.7 \times 10^{-7}$ in U-test; and for WRC: $p = 1.2 \times 10^{-3}$ in U-test). The occurrence of BCL6 breaks in BCL6-IG translocations at AID hotspot motifs but not CG motif indicates the germinal center origin of the arrangement.

The 53 BCL6 breakpoints that partner with non-IG loci in the 2156 bp region show significant proximity to CG motif ($p = 6.7 \times 10^{-4}$ in U-test in Table 2). Seventeen of them (32%) are right at CG ($p = 1.7 \times 10^{-3}$ in binomial test). The average distance of them to CG is 6.9 bp, versus 11.6 bp if they are at random ($p = 6.7 \times 10^{-4}$ in U-test). Twenty of the 53 breaks are right at AID WRC motif ($p = 9.1 \times 10^{-4}$ in binomial test). The average distance of them to WRC motif is 3.8 bp versus 4.8 bp if occurring at random ($p = 0.03$ in U-test). Six of the 53 BCL6 non-IG breaks in the fragile zone are within AID WGCW motif ($p = 0.026$ in binomial test), with an average distance of 25.5 bp versus 34.2 bp if randomly distributed ($p = 0.023$ in U-test). Overall, the BCL6 breakpoints in BCL6-non-IG translocations are highly focused on CG motif and show certain level of preference to AID hotspot motifs. The CG motif and AID hotspot motif seem to indicate the early B cell origin of the BCL6-non-IG translocations.

Among the 177 MYC breakpoints in IGH-MYC translocation, 156 (89%) of

them are randomly distributed within a 4.1 kb region between 1.4 kb upstream of *MYC* 5'UTR and its intron 1 without apparent clusters (Fig. S5B). The average distance of the MYC breakpoints to CG motif is 8.7 bp compared with 8.9 bp if they are in random pattern ($p = 0.77$ in U-test). CG motif does not seem to be a hotspot in MYC translocations. The MYC breakpoints in this 4.1 kb region do not show significant proximity to AID CGC motif ($p = 0.88$ in U-test). However, AID hotspot WGCW ($p = 6.7 \times 10^{-6}$ in U-test) and WRC ($p = 0.1$ in U-test) motifs are overrepresented around MYC breaks. The MYC breakpoints are neither close to CAC (RAG motif) in statistical analysis (data not shown). These results may suggest a later stage occurrence of the breakage in mature B cells caused by AID activity.

Local features around the narrow (20-600 bp) fragile zones

The CG motif and the AID hotspot motifs are very common in the human genome, averaging 1 per 80 in the human genome for CpG, and even more abundant for AID hotspot motifs. Other factors besides the local motif features must be involved to determine which regions have vulnerable CG or AID motifs. Reviewing a broader region surrounding fragile zones helps to check for distinctive features that might contribute to the fragility of the B cell fragile zones and may be in common among them the various 20 to 600 bp fragile zones of human B cell lymphomas.

Nucleotide composition and surrounding CG motif proximity

Considering that the CG motif is statistically important in the breakage phase of E2A and the other human fragile regions, we were wondering about the proximity of the nearest CpG sites upstream and downstream of the fragile regions (Table S1). The nearest CpG to the E2A fragile region is 97 bp upstream and 116 bp downstream.

Those two CpG sites define a region of about 240 bp in length where the CG motif only exists in the central 23 bp fragile region. This same point applies to the other fragile regions of < 600 bp. Though the break sites within the fragile regions at *MALT1*, *BCL1*, and *BCL2* are larger than the one on *E2A*, the broader potential ssDNA regions delimited by the nearby CG motif range between 200 bp to 1000 bp, with MALT1 fragile region and BCL1 MTC being close to the low end of this range (Table S1). For large fragile zones of several kilo base pairs in *MYC* and *BCL6*, the zone defined by the nearest CpG sites outside of the fragile regions is in similar size as to the original fragile regions.

DNA repeats

Repeated sequences in the genome could lead to transient misalignment of DNA during transcription, replication, local DNA repair synthesis, or any other processes that could separate two DNA strands. Misaligned regions will be in a ssDNA state transiently, and that state will be vulnerable to nucleases or AID inside the cells (Pannunzio and Lieber 2018). DNA direct repeats and inverted repeats with a length of 6-30 bp and the interspace between 0 bp and 30 bp in regions defined by the nearest CG sites outside of the fragile zones mentioned in Table S1 were checked.

BCL2 MBR contains three DNA direct repeats with the length of 6 bp, 7 bp, and 8 bp and an inverted DNA repeat of 6 bp (Fig. S10A). The 6 bp direct repeat (top panel of Fig. S10A, DR2, solid circles in grey) flanks the two CpG sites of the first MBR peak. The third peak of MBR is flanked by the 7 bp direct repeat (DR3, solid circles in yellow). A DNA repeat of 7 bp (top panel of Fig. S10A, DR4, solid circles in light blue) is right downstream of MBR. One direct repeat of 6 bp is located at the edge of icr of *BCL2*, overlapping with the second CpG site of icr (middle panel of Fig. S10A). A direct repeat of 6 bp and two inverted repeats of 6 bp and 7 bp are found right

downstream of the icr. Seven direct repeats and two inverted repeats are present in BCL2 mcr (bottom panel of Fig. S10A). A direct repeat of 7 bp (bottom panel of Fig. S10A, DR2, solid circles in yellow) is flanking the first CG motif of icr, where the 12 out of 19 breakpoints of mcr are clustered around.

Three direct DNA repeats and one inverted repeat are located within the 150 bp BCL1 MTC (Fig. S10B). The 6 bp inverted repeat (RR1, open circles in yellow) overlaps with the first CpG site in the fragile zone. The 6 bp direct repeat (DR3, solid circles in light blue) flanks the last two CpG sites in MTC.

A DNA direct repeat of 6 bp (Fig. S10C, DR2, shown in solid circles in light blue) in length flanks the 23 bp E2A fragile zone. An inverted DNA repeat of 6 bp (open circles in green, RR3) in and a direct DNA repeat of 7 bp (solid circles in dark blue, DR3) are found right downstream of the E2A fragile zone.

Both the 86 bp MALT1 fragile zone and the 311 bp CRLF2 fragile zone are in AT-rich regions. Increased numbers of DNA repeats are present in these two fragile zones due to the repetitive A- and T-strings. Six inverted repeats and two direct repeats are in the MALT1 fragile zone (Fig. S10D). The first CpG site within the MALT1 fragile zone, where most of the patients break in IGH-MALT1 translocation, is flanked by a 6 bp inverted repeat (Fig. S10D, RR1, open circles in brown). Five direct repeats and six inverted repeats are found within the 311 bp CRLF2 fragile zone (Fig. S10E). The first CpG site is flanked by a 6 bp direct repeat (Fig. S10E, DR2, solid circles in light blue). Several DNA repeats are located right upstream of the last two CpG sites within the CRLF2 break region, though none of them is flanking the two CpG sites.

Abundant DNA repeats are found in the large break regions of *BCL6* and *MYC* (data not shown). Eight-nine repeats, including 69 direct DNA repeats and 20 inverted DNA repeats are found in the 2156 bp BCL6 fragile region, averaging one repeat per 24

bp. A total of 73 direct repeats and 43 reverse repeats are found in the 4.1 kb MYC fragile zones (one repeat per 35 bp).

Except for the small DNA repeats (≤ 30 bp), there are many long repetitive regions in the genome such as LINE, SINE, LTR, and DNA, which occupy about 50% of human genome (Burns 2017). The long repetitive elements in a lymphoblastoid cell line (GM12878) were annotated with RepeatMasker in UCSC Genome Browser near all fragile zones. We found that long repetitive sequences are present in some but not all B cell fragile zones ("Repeat" panel in Fig. S9). Low complexity repeats are commonly observed in BCL6 and MYC fragile zones. Two LINE elements are located within the 175 bp MBR of *BCL2*. One LINE element and one SINE element are found within the 311 bp CRLF2 fragile zone. A transposable element MER20 sequence is located right downstream of E2A fragile zone. One study proposed that the significant enrichment of MER20 DNA transposon near the E2A fragile zone may be related to the E2A breakage (Rodic et al. 2013). We found the conclusion is unsolid because the MER20 sequence used for genome BLAST was truncated in that study which makes the statistical analyses bias for E2A fragile zone. The truncated MER20 near the E2A fragile zone misses all the lateral elements required for an active transposon. The invasion of MER20 near E2A fragile zone could be an occasional event resulted from DNA breakage rather the cause of the breakage event.

Histone modifications

Histone modifications can alter chromatin structure, affect the accessibility of the genes, impact gene expression, and serves as landmarks to recruit other proteins (Berger 2002; Latham and Dent 2007; Suganuma and Workman 2011; Zentner and Henikoff 2013; Lawrence et al. 2016). It has been intensively studied that histone modifications have a fundamental role in DNA DSB repair initiation and regulation in the rejoining phase

(Foster and Downs 2005; Van Attikum and Gasser 2009). However, studies regarding the effect of histone modifications on the breakage of DNA are limited. We hypothesize the histone modifications located specifically around the fragile regions may modify local DNA structure by itself or by recruiting other DNA structure modifiers. Once the local DNA structure has been changed, the DNA in certain region may become more accessible or be more prone to be in transient ssDNA state. It will increase the chance of the certain DNA sequence being targeted by nucleases and other enzymes that can lead to DSBs.

ChIP-seq results for CTCF and histone modifications including H2A.Z, H3K27ac, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H4K79me2, H3K27me3, H3K9ac, H3K9me3, and H4K20me1 near all fragile zones in a lymphoblastoid cell line (GM12878) are assembled from UCSC Genome Browser ("Histone modifications" panel in Fig. S9). H2A.Z, H3K79me2, and H3K36me3 are three neutral histone modifications that don't correlate with gene activation or repression. Histone variant H2A.Z usually locates at +1 and -1 nucleosomes to prevent the spread of the heterochromatin (Raisner et al. 2005; Barski et al. 2007). H3K79me2 and H3K36me3 are involved in many cellular processes including DNA repair, alternative splicing, and replication initiation (Kolasinska-Zwierz et al. 2009; Fu et al. 2013; Li F et al. 2013; Farooq et al. 2016; Li T et al. 2018). H3K27me3, H3K9me3, and H3K4me2 have been reported to mark the repression of genes (Barski et al. 2007; Schuettengruber et al. 2007; Liu Y et al. 2019). In contrast, H3K4me3, H3K9ac, H3K27ac, and H4K20me1 are related to gene activation and are associated with actively transcribed regions (Barski et al. 2007; Schuettengruber et al. 2007; Creyghton et al. 2010; Karmodiya et al. 2012; Liu X et al. 2016). We found that no single histone modification or a combination of different histone modifications are commonly present within all the

fragile zones. It indicates the breakage of the B cell fragile zones may not be caused by histone modifications or their associated biological processes.

A CTCF peak is present 200 bp upstream of the 23 bp E2A fragile zone in intron 16 (Fig. S9C). Three CTCF peaks are located within the MYC 4.1 kb break region (Fig. S9G). It is unclear if the CTCF binding has any effect on the breakage of those genes.

Transcription

Active transcription is an important contributor to transient ssDNA regions. Transcription separates the two DNA strands and leaves the non-transcribed strand (NTS) as ssDNA for the length (~15 bp) of the transcription bubble (Barnes et al. 2015). The template strand (TS) is paired with the nascent RNA transcript in the bubble. ssDNA, as we have been mentioned many times, is a potential substrate for AID and other nucleases.

The expression level of E2A, BCL2, BCL6, and MYC is very high in diffuse B cells which suggests an active transcription through these regions ("DNA features" panel in Fig. S9). LncRNA is another indicator of the transcription level of the fragile zones, especially for those located in the intergenic regions. We found the presence of lncRNA in the fragile zones of *E2A*, *BCL2*, *MALT1*, *BCL6*, and *MYC*, the latter two of which contribute to the situation of two conflicted promoters. Study regarding BCL6 fragile zone suggests the convergent BCL6 and lncRNA promoters may have a role in the BCL6 breakage (Lu, Pannunzio, et al. 2015).

Transient or intermittent pause of RNA polymerase during transcription can lead to prolonged duration of ssDNA in the transcription bubble. The probability of the NTS being attacked by enzymes inside the cells, such as by AID, can increase (Canugovi et al. 2009). We have found that DNA with C-strings (consecutive cytosines such as

CCCCC) is more reactive in native bisulfite reactions compared with DNA containing alternating guanines and cytosines (e.g. GCGCGC) (Tsai et al. 2009). The X-ray results indicate that the DNA duplex with C-strings adopt a non-canonical DNA structure that is intermediate between B form DNA and A form DNA (Dornberger et al. 1999). The intermediate B/A-DNA structure formed by C-strings could potentially delay the RNA polymerase II (pol II) during transcription. A recent study showed that human pol II has a frequent early transcription termination at C-string regions in the test tube (Pham et al. 2019). C-strings are found in B cell fragile zones including the 23 bp E2A fragile zone, MBR and icr of BCL2, BCL1 MTC, BCL6, and MYC (Table 2). Several assays were developed to map the pausing sites of RNA pol II in human genome ("TF binding" panel in Fig. S9). The genome-wide study of RNA pol II pausing sites in normal human skin fibroblast cells by Cheung et al. maps the consistent RNA pol II pausing sites across five individuals (Watts et al. 2019). Transient transcriptome sequencing (TT-seq) which measures the local rate of RNA synthesis and degradation, together with mNET-seq which shows the RNA pol II number detected within a certain window, enables the detection of the RNA pol II pausing duration at single nucleotide resolution in human Raji B cells (Schwalb et al. 2016; Gressel et al. 2017). None of these assays above indicate a strong RNA polII pausing sites within the fragile regions.

The ChIP-Seq results on RNA pol II in GM12878 cells were assembled from UCSC Genome Browser to investigate the abundance of pol II within the fragile regions ("TF binding" panel in Fig. S9). Several pol II peaks within BCL6 and MYC break regions are observed, with an absence of Pol II peaks within all other fragile zones. The low pol II signals in most fragile zones further indicate the low level of transcriptional pausing within these zones. The pol II signal observed in BCL6 and MYC fragile zones may indicate an increased transcription level or increased pausing rate. Considering the

low incidence of the breakage events, the pausing of RNA pol II could be occurring in very chance in fragile zones without pol II signal that is beyond the sensitivity of these assays.

Replication

DNA replication is another biological process which can separate two DNA strands and provide transient ssDNA for nuclease or deaminase action. Factors that can affect the replication fork could potentially contribute to DNA breakage.

The replication origins vary substantially in different cell types and under different conditions and methods used in different studies. A study using the HCT116 cell line showed that even with ORC2 knocked out, there are at least 52,000 replication origins being fired in each cell cycle (Shibata et al. 2016). Those replication origins do not seem to have any sequence preference. Another study using the HeLa cell line claims that origin density is associated with CpG islands based on the 283 origins identified in their investigation (Cadoret et al. 2008). A more recent genome-wide study shows that many of the DNA replication origin locations are conserved across 4 cell lines (IMR-90, HeLa, hESC H9, and iPSCs from IMR-90) (Picard et al. 2014). A G-quadruplex–forming DNA motif was identified as a conserved motif that can predict replication origins of human cells, but that has not been further pursued to our knowledge (Besnard et al. 2012; Langley et al. 2016).

Repli-Seq is newly developed assay to map the sequences of nascent DNA replication strands throughout the whole genome during each of the six cell cycle phases, having been used to genome-wide assessment of how cellular processes are linked to replication timing (Marchal et al. 2018). Repli-Seq results from the GM12878 cell line were assembled from UCSC Genome Browser ("Replication origins" panel in Fig. S9). All fragile zones are replicated at certain levels in various phases without

preference to any single phase. It has been reported that AID has increased stability and enhanced nuclear localization in G1 phase compared with other phases which restricted its activity in G1 phase (Le and Maizels 2015; Wang et al. 2017). Some level of replication was observed in G1 phase which generates ssDNA substrate for AID, potentially in all fragile zones except BCL2 MBR and BCL1 MTC.

Accessibility of the fragile zones

The accessibility of the fragile region is a key determinant of whether AID and other nucleases can obtain physical access to that portion of the DNA duplex. DNase I hypersensitivity (DNase I HS) assays are widely used to map the regions that are sensitive to DNase I cleavage (Wu 1980). The sequencing peaks from DNase I HS assay represent regions that have lost the condensed structure and are therefore in a more exposed state. Formaldehyde-assisted isolation of regulatory elements with deep sequencing (FAIRE-Seq) is a robust assay for detection of regulatory element binding regions, the signal of which indicates the accessibility of certain region (Giresi et al. 2007). Micrococcal nuclease digestion with deep sequencing (MNase-Seq) was developed to measure the nucleosome occupancy in the human genome (Schones et al. 2008). The occupied DNA segments are sequenced in this assay and displayed as peaks after mapping back to human reference genome. DNase I HS assay, FAIRE-Seq, and MNase-Seq are three methods that can provide strong cross-validation with each other.

We checked the DNase I HS assay, FAIRE-Seq, and MNase-Seq results for all the fragile zones in GM12878 cell line ("Chromatin accessibility" panel in Fig. S9). The BCL2 MBR is highly inaccessible based on the low signals in DNase I HS and FAIRE-Seq assays and obvious occupancy by nucleosomes (Fig. S9A). The BCL2 icr is in the middle of a DNase I peak and a FAIRE-Seq peak with low nucleosome occupancy, all of which indicates the high accessibility of this region. The broad region

covered by the 561 bp BCL2 mcr contains two nucleosome peaks at the edges with largely low nucleosome occupancy in the middle of this zone. DNase I signals are not observed in mcr but a moderate FAIRE-Seq peak is located within this region which may suggest the moderate accessible state of mcr. The 150 bp BCL1 MTC shows moderate DNase I signal and FAIRE-Seq signal together with a relative strong nucleosome occupancy peak which makes it difficult to draw a solid conclusion on the accessibility of MTC (Fig. S9B). The 23 bp E2A fragile region is absent of nucleosome signal and within a moderate FAIRE-Seq peak, suggesting an open state of this region though the DNase I HS assay only shows background signals (Fig. S9C). The 86 bp MALT1 fragile zone is in a nucleosome free region with moderate FAIRE-Seq signals (Fig. S9D). The 311 bp CRLF2 fragile zone contains an obvious FAIRE-Seq peak, suggesting a relative accessible state of this region (Fig. S9E). The 2156 BCL6 break region contains moderate FAIRE-Seq signal and low DNase I signal, with several peaks for nucleosome occupancy (Fig. S9F), which may indicate a relatively closed state. In contrast, MYC break region shows high DNase I signals and FAIRE-Seq signals, suggesting a very accessible state (Fig. S9G).

Though not always consistent with each other, the three assays indicate a largely open state of all fragile zones in B cells except BCL2 MBR (Fig. S9) (Lu, Lieber, et al. 2015). The general accessible state of the fragile zones indicates that the fragility of these region in human genome may correlate with the regional accessibility. The low accessibility of BCL2 MBR may suggest the involvement of other factors contributing to the transient ssDNA state of those fragile regions, which make it vulnerable to DSBs. One possibility is that the BCL2 MBR, occupied by nucleosomes, may be under distortion in living cells. Within each nucleosome, 146 bp of DNA is tightly wrapped around the histone octamer inside the cells. The high-resolution X-ray structure of the

nucleosome core particle shows that the superhelix around the histone octamer is not uniformly bent but has sharp curvatures and local kinks (Luger et al. 1997). Each nucleotide base pair suffers from varied levels of lateral shearing. The degree and location of DNA deformation inside the cells is decided by multiple factors including DNA sequence, histone modifications, and DNA binding proteins (Gasser 2016). The local distortion in MBR might be further increased by other factors and finally lead to transient ssDNA of the fragile region.

Cytosine methylation

U:G mismatches are repaired very quickly in mammalian cells by uracil DNA glycosylase (UDG). Thymine DNA glycosylase (TDG) and MBD4 are more than 2000-fold less efficient compared with UDG when excising the T:G mismatches (Schmutte et al. 1995; Walsh and Xu 2006). Therefore, the T:G mismatch resulting from AID deaminated methylcytidine is a long-lived lesion in the human genome. The persistent DNA lesions are more likely to be converted to DSBs inside the cells.

The results from reduced representation bisulfite sequencing (RRBS) and methylated DNA immunoprecipitation sequencing (MeDIP-Seq) are assembled from UCSC Genome Browser for all fragile zones ("DNA features" panel in Fig. S9). The cytosines within BCL2, BCL1, E2A, MALT1, and CRLF3 fragile zones have certain level of methylation in GM12878 cell line. These results support the possible events of persistent T:G mismatches in the fragile zones, once the methylated cytosines are deaminated by AID. For BCL6 and MYC break regions in mature B cells, CpG islands are founded, which are not present within or near fragile zones of early B cells. Interestingly, hypomethylated regions are observed in BCL6 and MYC break region together with increased nucleotide mutation rates (Fig. S9F; Fig. S9G). The different methylation pattern of fragile regions of early B cells and that of mature B cells may be

involved in different breakage mechanisms.

Supplemental Figure and Table legends

Figure S1. Mechanism of V(D)J recombination. V(D)J recombination occurs at sequences called 12-recombination signal sequence (RSS) and 23-RSS (see the figure). An RSS contains conserved heptamer and nonamer sequence elements, separated by either 12 or 23 non-conserved base pairs, hence the designation 12-RSS and 23-RSS. One recombination event requires one 12-RSS and one 23-RSS, and this is called the '12/23 rule'. Along with the constitutively expressed high mobility group box 1 (HMGB1) protein, the early lymphoid-specific recombination activating gene 1 (RAG1) and RAG2 proteins form a complex, designated the RAG complex, which nicks and then hairpins the DNA ends at the V and J segments (see the figure). The Ku complex (comprised of KU70-KU80) can bind to any of the four DNA ends. The DNA-PKcs complex then binds to the V and J hairpin ends and nicks the hairpins in a manner that usually results in a 3' overhang. The Artemis:DNA-PKcs complex can then further endonucleolytically resect at any 3' or 5' overhang. DNA pol μ and pol λ can fill-in the gap in a template-independent manner. The ligase complex includes XLF (also known as Cernunnos), XRCC4, and DNA ligase IV. Some antigen receptor loci have not only V and J segments, but also D segments; hence, the name V(D)J recombination. TdT is also expressed in early T and B cells, and it is responsible for most of the junctional addition of nucleotides, which is the major factor in what is called junctional diversity.

Figure S2. Mechanism of mammalian Ig heavy locus class switch recombination. (A) Mammalian *IGH* switch (*IGH*-S) regions form R-loops upon transcription, and these have now been demonstrated to be important for the efficiency of *IGH* CSR (Zhang, Pannunzio, Han, et al. 2014; Zhang, Pannunzio, Hsieh, et al. 2014). (B) AID requires ssDNA in order to recognize cytosine (shown as C in the figure) as a substrate, and R-loops provide a fully single-stranded NTS. AID has a preference for C that is surrounded by the sequence WGCW, where W = A or T. But AID can deaminate any C to a U (and any methylated C to a T, though at an efficiency that is somewhat lower) (Bransteitter et al. 2003). (C) Once AID has converted some of the Cs in the switch region to U, then UDG (also known as UNG, specifically UNG2) can remove the U to create an abasic site (dashed line in figure). Next, APE1 can create a nick at the abasic

site. This explains nicks on the NTS (Masani et al. 2013). (D) RNase H can remove portions of the RNA (RNA is shown as red line in the figure) that are annealed to the TS, thereby exposing ssDNA regions – thus allowing AID/UDG/APE action there as well (depicted as gap in the figure). AID can also act on the TS at the edge of the R-loop (Yu et al. 2005). (E) NHEJ is the primary pathway for joining the DNA ends (Han and Yu 2008).

Figure S3. Mechanistic aspects of Ig somatic hypermutation. Although not fully understood, some of the known elements of the SHM process are shown in the figure. AID requires ssDNA (Bransteitter et al. 2003; Pham et al. 2003). Unlike for *IGH* CSR, where stable kilobase length R-loops provide ssDNA, for SHM at physiological loci (primarily *IGH* and *IGL* V segments), the ssDNA may arise simply due to transcription, which is accompanied by transient underwinding (negative superhelical tension) of the dsDNA in the wake of the RNA polymerase as it passes through a region. After AID-mediated deamination of C to U on either strand of the DNA (or methylated C to T), then either of two error-free mechanisms may repair the site, without mutation. These two error-free mechanisms are base excision repair (BER) and mismatch repair (MMR). But in SHM, any of three error-prone pathways may operate. First, an error-free DNA polymerase may simply copy the U, which is read as a T, resulting in a C:G to T:A transition. Second, UDG will remove the U, and APE1 will nick 5' to the abasic site. This will then allow a REV1-dependent transversion to result. Third, after UDG and APE1 action, exonuclease 1 (EXO1) can resect downstream of the nick site, providing a long gap that is filled-in by the error-prone DNA polymerase eta (pol **η**).

Figure S4. Schematic illustration of B cell translocations. (A) Schematic illustration of BCL2-IGH translocation. The breakage of *IGH* locus is caused by RAG complex during V(D)J recombination at the D_H and J_H segments. A mismatch on *BCL2* is initiated by AID and then recognized by Artemis:DNA-PKcs complex to further be converted to a DSB. The DNA ends from breaks at the IGH loci are mistakenly joined with ends from BCL2 breaks. This results in the *BCL2* gene being under the regulation of the enhancer region of IGH (Eμ). (B) Schematic illustration of the CCND1-IGH translocation. The CCND1-IGH translocations occur in a similar way as the BCL2-IGH translocation. The expression of *CCND1* gene is regulated by the enhancer region of IGH after the translocation event. (C) Schematic illustration of E2A-PBX1

translocation. The breakage of *PBX1* gene is mainly restricted to its intron 2 and the *E2A* gene breakpoints are localized in intron 16. The E2A exon 1 to exon 16 is fused to *PBX1* (exons 3 to 9), resulting in a chimeric protein containing the E2A transactivation domain and PBX1 DNA binding domain. (D) Schematic illustration of the BCL6-IGH translocation. In this translocation, the breakpoints of the IGH locus are mainly restricted to the switch region. The BCL6 breakpoints span from the upstream region of the gene to its intron 1. Exon 1 of *BCL6* is non-coding. The translocation event fuses part of the switch region to the coding region of *BCL6* (starting at exon 2), resulting in dysregulation of *BCL6* expression. (E) Schematic illustration of the MYC-IGH translocation. The breakpoints of *MYC* span from the upstream region of the *MYC* gene to its intron 1. Exon 1 of *MYC* is non-coding. The translocation event leads to *MYC* under the regulation of IGH enhancer. E and I: enhancer; C: constant region; S: switch region; V, D, and J: variable, diversity, and junction; E: exon of a gene.

Figure S5. Breakpoint distribution on *BCL6* and *MYC* genes. (A) BCL6 breakpoint distribution. The BCL6 breakpoints span over an 18 kb region from 6.7 kb upstream of *BCL6* to its intron 2. Around 88% of the BCL6 breakpoints are clustered within a 2156 bp region in *BCL6* intron 1. The BCL6 breakpoints in BCL6-IG translocations (IG type) are shown in the top panel. The breakpoints located outside of the 2156 bp BCL6 break region in IG type translocations tend to distribute toward the telomeric side of the gene. The BCL6 breakpoints in translocations involving non-IG loci (non-IG type) are shown in the bottom panel and tend to distribute toward the centromeric side of the gene. (B) MYC breakpoints distribution. Over 88% of MYC breakpoints in MYC-IGH translocations are clustered in a 4.1 kb region covering the upstream zone of *MYC* to its intron 1. The MYC break region contains high density of CG motif. The CG motif in both figures is highlighted in the red background.

Figure S6. Three pathways that lead to DSBs mediated by AID. AID can deaminate methylcytosine in single stranded region at CG motif to form a long-lived lesion. A second deamination event could occur on the opposite strand, and a 2 bp bubble structure is formed. Within the small bubble (heterologous loop), the methylcytosine can be recognized by methyl-CpG binding domain 4 (MBD4) or thymine DNA glycosylase (TDG) to be converted to an abasic site, which can be further cut by APE1 nuclease. The structure specific nucleases inside the cells such as RAG complex and Artemis:DNA-PKcs complex can nick at the mismatches to create DSBs [the red arrows

in the second line are possible nucleolytic cut sites by either Artemis:DNA-PKcs or by RAG1/2 (Tsai et al. 2008; Cui et al. 2013)].

Figure S7. Distribution of PBX1 and HLF breakpoints. (A) PBX1 breakpoint distribution. The PBX1 breakpoints in 48 out 49 patients are randomly dispersed in its intron 2 with no obvious clustering. (B) HLF breakpoint distribution. HLF breakpoints in 7 out of 8 patients locate within its intron 3 with no obvious motif proximity or clustering.

Figure S8. Distribution of C-strings around the fragile zones involved in B cell translocations. C-strings distribution around (A) the 175 bp MBR, 105 bp icr, and 561 bp mcr of *BCL2*, (B) the 150 bp BCL1 (CCND1) MTC, (C) the 23 bp E2A fragile zone, (D) the 89 bp MALT1 fragile zone, (E) the 311 bp CRLF2 fragile zone, (F) the 2156 bp BCL6 break region, and (G) the 4.1 kb MYC break region. For all panels: the x axis denotes each position within the 3 kb to 10 kb region around the fragile zones. The y axis shows the length of the C-string with the positive direction for the non-template strand and the negative direction for the template strand (template strands are defined in accord with the genes for which the fragile zones are named). Only C-strings with a length of four or more are shown in all the figures. The two dashed vertical red lines in each figure define the boundaries of the fragile zones.

Figure S9. Regional features around the fragile zones of early and mature B cells. Regional features around BCL2 (A), BCL1(CCND1) (B), E2A (C), MALT1 (D), CRLF2 (E), BCL6 (F), and MYC (G) break regions. A combination of features around the regions containing the fragile zones was assembled from UCSC Genome Browser to investigate the potential factors that could contribute to the fragility of all fragile zones of early B cells. All features are from the GM12878 cell line if not specifically annotated below. The coordinates of the regions from hg19 are shown at the top along with the scale.

The first panel (DNA features) contains information for lncRNA, expression level in diffuse large B cells, mutation sites in malignant lymphomas, CpG methylation pattern [resulting from both reduced representation bisulfite sequencing (RRBS) and methylated DNA immunoprecipitation sequencing (MeDIP-Seq)], and the distribution of CpG islands. A darker color for the expression level in diffuse large B cells represents a higher expression level. For RRBS, the red indicates 100% of molecules

sequenced are methylated; yellow means 50% of molecules sequenced are methylated; green represents 0% of molecules sequenced are methylated. The methylation level in MeDIP-Seq is increased as with the increase of the darkness of the vertical bars. CpG islands shown fulfill the criteria listed below: 1. GC content of 50% or greater; 2. length greater than 200 bp; 3. ratio greater than 0.6 of observed number of CG dinucleotides to the expected number on the basis of the number of Gs and Cs in the segment. The increased size of the CpG island is illustrated with a darker green color.

The second chromatin accessibility panel shows results from DNase I hypersensitivity assay (DNase I HS), formaldehyde-assisted isolation of regulatory elements followed by sequencing (FAIRE-Seq), and micrococcal nuclease digestion followed by sequencing (MNase-Seq). The peaks shown in DNase I HS represents the accessible regions by DNase I nuclease. The FAIRE-Seq peaks indicate regions not bound by nucleosomes and proteins. The MNase-Seq method sequences DNA bound by nucleosomes and the peaks represent regions occupied by nucleosomes.

The third panel shows the binding of different transcription factors (TFs). Chromatin immunoprecipitation sequencing (ChIP-Seq) results for CTCF and RNA polymerase II (Pol II) are listed on top. The strand-specific results of mNET-seq and TT-seq in Raji B cell line are illustrated at the bottom with two replicates for each assay. Those two assays measure the nascent transcript and the number of RNA polymerase II within a sequence window which can provide information on potential RNA polymerase II pausing sites (Schwalb et al. 2016; Gressel et al. 2017). The y axis is fixed between 0 and 250 for TT-seq and between 0 and 150 for mNET-seq for the convenience of comparison between strand-specific tracks.

The fourth panel shows ChIP-Seq results for different histone modifications including H2A.Z, H3K27ac, H3K36me3, H3K4me1, H3K4me2, H3K4me3, H4K79me2, H3K27me3, H3K9ac, H3K9me3, and H4K20me1. The y-axis is in auto scale in order to show all existing peaks around the fragile zones.

The fifth panel presents the results from Repli-Seq that maps the sequences of nascent DNA replication strands throughout the whole genome during each of the six cell cycle phases. The higher replication frequency is indicated with a darker color.

The sixth panel shows the long repetitive DNA sequences including SINE, LINE, LTR, etc. around the fragile zones.

Figure S10. Distribution of DNA repeats near the fragile zones of early B cells. The DNA repeats with the length of 6 bp to 30 bp and the interspace of 0 bp to 30 bp were screened within regions defined by the nearest CG motif outside of each fragile zone as shown in Table S1. The fragile zones are indicated in each figure by solid black circles aligned with the x axis. A pair of DNA repeats is indicated by dots of matching color in all figures. The circles were plotted according to the starting and ending positions and length of the DNA repeats. The direct repeats are shown in solid circles and the inverted ones are shown in open circles. The 323 bp region around BCL2 MBR contains 6 direct repeats and 2 inverted repeats (top panel in A). The 518 bp region surround BCL2 icr has 4 direct repeats and 7 inverted repeats (middle panel in A). There are 10 direct repeats and 10 inverted repeats in the 1067 bp region surrounding BCL2 mcr (bottom panel in A). The 295 bp region surrounding BCL1 MTC contains 8 direct repeats and 1 inverted repeat (B). Three direct repeats and 5 inverted repeats are found in the 236 bp region surrounding the 23 bp E2A fragile zone (C). The 280 bp zone around the 89 bp MALT1 fragile region contains 2 direct repeats and 7 inverted repeats (D). There are 14 direct repeats and 7 inverted repeats in the 491 bp region surrounding the 311 bp CRLF2 fragile zone (E).

Figure S11. Key factors for DNA breakage in 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. A 236 bp region where only the 23 bp fragile zone contains CG is shown as green horizontal line between the two exons. The E2A fragile zone is marked by the red asterisks above the black line. 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 bracket. 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. The two CpG sites within the E2A fragile zone are both in WRCG motifs. They are the only two WRCG motif that overlap with each other within the 666 bp region. (B) Illustration of the WRCG sites within the 23 bp E2A fragile zone. Sequence around the 23 bp E2A

fragile zone is illustrated in this figure. The E2A fragile zone is shown in red nucleotides with the CG motif in green. The two WRCG sites are circled in blue boxes.

Table S1. Regions defined by the nearest CG motif outside of the fragile zones. The distance of the nearest CG motif located upstream and downstream of each fragile zone is listed in the first and third column. The size of the region defined by the nearest CpG sites located outside of the fragile regions are shown in the fourth column. The upstream and downstream CpG sites define the length of regions for which the only CG motif is located within the break regions.

References

Akasaka H, Akasaka T, Kurata M, Ueda C, Shimizu A, Uchiyama T, Ohno H. 2000. Molecular anatomy of BCL6 translocations revealed by long-distance polymerase chain reaction-based assays. Cancer Research. 60(9):2335-2341.

Bain G, Maandag ECR, Izon DJ, Amsen D, Kruisbeek AM, Weintraub BC, Krop I, Schlissel MS, Feeney AJ, van Roon M. 1994. E2A proteins are required for proper B cell development and initiation of immunoglobulin gene rearrangements. Cell. 79(5):885-892.

Barnes CO, Calero M, Malik I, Graham BW, Spahr H, Lin G, Cohen AE, Brown IS, Zhang Q, Pullara F et al. 2015. Crystal Structure of a Transcribing RNA Polymerase II Complex Reveals a Complete Transcription Bubble. Molecular Cell. 59(2):258-269.

Barski A, Cuddapah S, Cui K, Roh T-T, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K. 2007. High-Resolution Profiling of Histone Methylations in the Human Genome. Cell. 129(4):823-837.

Basso K, Dalla-Favera R. 2012. Roles of BCL6 in normal and transformed germinal center B cells. Immunological Reviews. 247(1):172-183.

Bastard C, Tilly H, Lenormand B, Bigorgne C, Boulet D, Kunlin A, Monconduit M, Piguet H. 1992. Translocations Involving Band 3q27 and Ig Gene Regions in Non-Hodgkin's Lymphoma. Blood. 79(10):2527-2531.

Battey J, Moulding C, Taub R, Murphy W, Stewart T, Potter H, Lenoir G, Leder P. 1983. The human c-myc oncogene: structural consequences of translocation into the IgH locus in Burkitt lymphoma. Cell. 34(3):779-787.

Berger SL. 2002. Histone modifications in transcriptional regulation. Current opinion in genetics & development. 12(2):142-148.

Besnard E, Babled A, Lapasset L, Milhavet O, Parrinello H, Dantec C, Marin J-M, Lemaitre J-M. 2012. Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nature Structural & Molecular Biology. 19(8):837-844.

Boxer LM, Dang CV. 2001. Translocations involving c-myc and c-myc function. Oncogene. 20(40):5595-5610.

Bransteitter R, Pham P, Scharff MD, Goodman MF. 2003. Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. Proc Natl Acad Sci U S A. 100(7):4102-4107.

Buchonnet G, Lenain P, Ruminy P, Lepretre S, Stamatoullas A, Parmentier F, Jardin F, Duval C, Tilly H, Bastard C. 2000. Characterisation of BCL2-JH rearrangements in

follicular lymphoma: PCR detection of 3′ BCL2 breakpoints and evidence of a new cluster. Leukemia. 14(9):1563-1569.
Burns KH. 2017. Transposable elements in cancer. Nature Reviews Cancer. 17(7):415-424.
Cadoret J-C, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, Duret L, Quesneville H, Prioleau M-N. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proceedings of the National Academy of Sciences. 105(41):15837-15842.
Campos L, Rouault J-P, Sabido O, Oriol P, Roubi N, Vasselon C, Archimbaud E, Magaud J-P, Guyotat D. 1993. High expression of bcl-2 protein in acute myeloid leukemia cells is associated with poor response to chemotherapy.
Canugovi C, Samaranayake M, Bhagwat AS. 2009. Transcriptional pausing and stalling causes multiple clustered mutations by human activation-induced deaminase. The FASEB Journal. 23(1):34-44.
Carvajal-Garcia J, Cho J-E, Carvajal-Garcia P, Feng W, Wood RD, Sekelsky J, Gupta GP, Roberts SA, Ramsden DA. 2020. Mechanistic basis for microhomology identification and genome scarring by polymerase theta. Proceedings of the National Academy of Sciences. 117(15):8476-8485.
Chang HHY, Pannunzio NR, Adachi N, Lieber MR. 2017. Non-homologous DNA end joining and alternative pathways to double-strand break repair. Nat Rev Mol Cell Bio. 18(8):495.
Chen Y-W, Hu X-T, Liang AC, Au W-Y, So C-C, Wong ML, Shen L, Tao Q, Chu K-M, Kwong Y-L. 2006. High BCL6 expression predicts better prognosis, independent of BCL6 translocation status, translocation partner, or BCL6-deregulating mutations, in gastric lymphoma. Blood. 108(7):2373-2383.
Cleary ML, Smith SD, Sklar J. 1986. Cloning and structural analysis of cDNAs for bcl-2 and a hybrid bcl-2/immunoglobulin transcript resulting from the t (14; 18) translocation. Cell. 47(1):19-28.
Cory S. 1986. Activation of cellular oncogenes in hemopoietic cells by chromosome translocation. Advances in cancer research. 47:189-234.
Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA et al. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences. 107(50):21931-21936.
Cui X, Lu Z, Kurosawa A, Klemm L, Bagshaw AT, Tsai AG, Gemmell N, Müschen M, Adachi N, Hsieh CL et al. 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(5):947-957.
Deweindt C, Kerckaert JP, Tilly H, Quief S, Nguyen VC, Bastard C. 1993. Cloning of a breakpoint cluster region at band 3q27 involved in human non-Hodgkin's lymphoma. Genes Chromosomes Cancer. 8(3):149-154.
Dornberger U, Leijon M, Fritzsche H. 1999. High base pair opening rates in tracts of GC base pairs. J Biol Chem. 274(11):6957-6962.
Farooq Z, Banday S, Pandita TK, Altaf M. 2016. The many faces of histone H3K79 methylation. Mutation Research/Reviews in Mutation Research. 768:46-52.
Finnon P, Lloyd DC, Edwards AA. 1995. Fluorescence in situ hybridization detection of chromosomal aberrations in human lymphocytes: applicability to biological dosimetry. International Journal of Radiation Biology. 68(4):429-435.

Fischer U, Forster M, Rinaldi A, Risch T, Sungalee S, Warnatz H-J, Bornhauser B, Gombert M, Kratsch C, Stütz AM. 2015. Genomics and drug profiling of fatal TCF3-HLF− positive acute lymphoblastic leukemia identifies recurrent mutation patterns and therapeutic options. Nature genetics. 47(9):1020-1029.
Foa R, Vitale A, Mancini M, Cuneo A, Mecucci C, Elia L, Lombardo R, Saglio G, Torelli G, Annino L. 2003. E2A–PBX1 fusion in adult acute lymphoblastic leukaemia: biological and clinical features. British journal of haematology. 120(3):484-487.
Foster ER, Downs JA. 2005. Histone H2A phosphorylation in DNA double-strand break repair. The FEBS Journal. 272(13):3231-3240.
Fu H, Maunakea AK, Martin MM, Huang L, Zhang Y, Ryan M, Kim R, Lin CM, Zhao K, Aladjem MI. 2013. Methylation of histone H3 on lysine 79 associates with a group of replication origins and helps limit DNA replication once per cell cycle. PLoS Genet. 9(6):e1003542.
Gasser SM. 2016. Nuclear Architecture: Past and Future Tense. Trends in Cell Biology. 26(7):473-475.
Gauss GH, Lieber MR. 1996. Mechanistic constraints on diversity in human V(D)J recombination. Mol Cell Biol. 16(1):258-269.
Giresi PG, Kim J, McDaniell RM, Iyer VR, Lieb JD. 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Research. 17(6):877-885.
Greisman HA, Lu Z, Tsai AG, Greiner TC, Yi HS, 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(14):2864-2867.
Gressel S, Schwalb B, Decker TM, Qin W, Leonhardt H, Eick D, Cramer P. 2017. CDK9-dependent RNA polymerase II pausing controls transcription initiation. Elife. 6:e29736.
Han L, Yu K. 2008. Altered kinetics of nonhomologous end joining and class switch recombination in ligase IV–deficient B cells. Journal of Experimental Medicine. 205(12):2745-2753.
Harvey RC, Mullighan CG, Chen IM, Wharton W, Mikhail FM, Carroll AJ, Kang H, Liu W, Dobbin KK, Smith MA et al. 2010. Rearrangement of CRLF2 is associated with mutation of JAK kinases, alteration of IKZF1, Hispanic/Latino ethnicity, and a poor outcome in pediatric B-progenitor acute lymphoblastic leukemia. Blood. 115(26):5312-5321.
Hayday AC, Gillies SD, Saito H, Wood C, Wiman K, Hayward WS, Tonegawa S. 1984. Activation of a translocated human c-myc gene by an enhancer in the immunoglobulin heavy-chain locus. Nature. 307(5949):334-340.
Hecht JL, Aster JC. 2000. Molecular biology of Burkitt's lymphoma. Journal of Clinical Oncology. 18(21):3707-3721.
Hein D, Dreisig K, Metzler M, Izraeli S, Schmiegelow K, Borkhardt A, Fischer U. 2019. The preleukemic TCF3-PBX1 gene fusion can be generated in utero and is present in≈ 0.6% of healthy newborns. Blood. 134(16):1355-1358.
Herold T, Schneider S, Metzeler KH, Neumann M, Hartmann L, Roberts KG, Konstandin NP, Greif PA, Bräundl K, Ksienzyk B et al. 2017. Adults with Philadelphia chromosome-like acute lymphoblastic leukemia frequently have IGH-CRLF2 and JAK2 mutations, persistence of minimal residual disease and poor prognosis. Haematologica. 102(1):130-138.
Hertzberg L, Vendramini E, Ganmore I, Cazzaniga G, Schmitz M, Chalker J, Shiloh R, Iacobucci I, Shochat C, Zeligson S. 2010. Down syndrome acute lymphoblastic leukemia, a highly heterogeneous disease in which aberrant expression of CRLF2 is

associated with mutated JAK2: a report from the International BFM Study Group. Blood, The Journal of the American Society of Hematology. 115(5):1006-1017.
Hunger SP, Ohyashiki K, Toyama K, 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. Gene Dev. 6(9):1608-1620.
Inaba T, Roberts WM, Shapiro LH, Jolly KW, Raimondi SC, Smith SD, Look AT. 1992. Fusion of the leucine zipper gene HLF to the E2A gene in human acute B-lineage leukemia. Science. 257(5069):531-534.
Jager U, Bocskor S, Le T, Mitterbauer G, Bolz I, Chott A, Kneba M, Mannhalter C, 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(11):3520-3529.
Jain N, Lu X, Daver N, Thakral B, Wang SA, Konoplev S, Patel K, Kanagal-Shamanna R, Valentine M, Tang G et al. 2017. Co-occurrence of CRLF2-rearranged and Ph+ acute lymphoblastic leukemia: a report of four patients. Haematologica. 102(12):e514-e517.
Kamps MP, Look AT, Baltimore D. 1991. The human t(1;19) translocation in pre-B ALL produces multiple nuclear E2A-Pbx1 fusion proteins with differing transforming potentials. Genes Dev. 5(3):358-368.
Karmodiya K, Krebs AR, Oulad-Abdelghani M, Kimura H, Tora L. 2012. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. BMC Genomics. 13(1):424.
Kato M, Ishimaru S, Seki M, Yoshida K, Shiraishi Y, Chiba K, Kakiuchi N, Sato Y, Ueno H, Tanaka H. 2017. Long-term outcome of 6-month maintenance chemotherapy for acute lymphoblastic leukemia in children. Leukemia. 31(3):580-584.
Kee BL, Quong MW, Murre C. 2000. E2A proteins: Essential regulators at multiple stages of B-cell development. Immunological reviews. 175(1):138-149.
Kitano M, Moriyama S, Ando Y, Hikida M, Mori Y, Kurosaki T, Okada T. 2011. Bcl6 Protein Expression Shapes Pre-Germinal Center B Cell Dynamics and Follicular Helper T Cell Heterogeneity. Immunity. 34(6):961-972.
Knoepfler PS, Kamps MP. 1995. The pentapeptide motif of Hox proteins is required for cooperative DNA binding with Pbx1, physically contacts Pbx1, and enhances DNA binding by Pbx1. Mol Cell Biol. 15(10):5811-5819.
Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. 2009. Differential chromatin marking of introns and expressed exons by H3K36me3. Nature genetics. 41(3):376.
Langley AR, Gräf S, Smith JC, Krude T. 2016. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). Nucleic Acids Research. 44(21):10230-10247.
Latham JA, Dent SYR. 2007. Cross-regulation of histone modifications. Nature structural & molecular biology. 14(11):1017-1024.
Lawrence M, Daujat S, Schneider R. 2016. Lateral thinking: how histone modifications regulate gene expression. Trends in Genetics. 32(1):42-56.
Le Q, Maizels N. 2015. Cell Cycle Regulates Nuclear Stability of AID and Determines the Cellular Response to AID. PLoS Genet. 11(9):e1005411.
LeBrun DP, Cleary ML. 1994. Fusion with E2A alters the transcriptional properties of the homeodomain protein PBX1 in t(1;19) leukemias. Oncogene. 9(6):1641-1647.

Li F, Mao G, Tong D, Huang J, Gu L, Yang W, Li G-M. 2013. The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. Cell. 153(3):590-600.
Li T, Liu Q, Garza N, Kornblau S, Jin VX. 2018. Integrative analysis reveals functional and regulatory roles of H3K79me2 in mediating alternative splicing. Genome Medicine. 10(1):30.
Lieber MR. 2016. Mechanisms of human lymphoid chromosomal translocations. Nat Rev Cancer. 16(6):387-398.
Lin CH, Wang Z, Duque-Afonso J, Wong SH, Demeter J, Loktev AV, Somervaille TCP, Jackson PK, Cleary ML. 2019. Oligomeric self-association contributes to E2A-PBX1-mediated oncogenesis. Sci Rep. 9(1):4915.
Liu X, Wang C, Liu W, Li J, Li C, Kou X, Chen J, Zhao Y, Gao H, Wang H et al. 2016. Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos. Nature. 537(7621):558-562.
Liu Y, Liu K, Yin L, Yu Y, Qi J, Shen W-H, Zhu J, Zhang Y, Dong A. 2019. H3K4me2 functions as a repressive epigenetic mark in plants. Epigenet Chromatin. 12(1):40.
Lu Z, Lieber MR, Tsai AG, Pardo CE, Müschen M, Kladde MP, Hsieh CL. 2015. Human lymphoid translocation fragile zones are hypomethylated and have accessible chromatin. Mol Cell Biol. 35(7):1209-1222.
Lu Z, Pannunzio NR, Greisman HA, Casero D, Parekh C, Lieber MR. 2015. Convergent BCL6 and lncRNA promoters demarcate the major breakpoint region for BCL6 translocations. Blood, The Journal of the American Society of Hematology. 126(14):1730-1731.
Lu Z, Tsai AG, Akasaka T, Ohno H, Jiang Y, Melnick AM, Greisman HA, Lieber MR. 2013. BCL6 breaks occur at different AID sequence motifs in Ig–BCL6 and non-Ig–BCL6 rearrangements. Blood. 121(22):4551-4554.
Lucas PC, Yonezumi M, Inohara N, McAllister-Lucas LM, Abazeed ME, Chen FF, Yamaoka S, Seto M, Núñez G. 2001. Bcl10 and MALT1, independent targets of chromosomal translocation in malt lymphoma, cooperate in a novel NF-κB signaling pathway. J Biol Chem. 276(22):19012-19019.
Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. 1997. Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature. 389(6648):251-260.
Maga G, Hübscher U. 2003. Proliferating cell nuclear antigen (PCNA): a dancer with many partners. J Cell Sci. 116(15):3051-3060.
Marchal C, Sasaki T, Vera D, Wilson K, Sima J, Rivera-Mulia JC, Trevilla-García C, Nogues C, Nafie E, Gilbert DM. 2018. Genome-wide analysis of replication timing by next-generation sequencing with E/L Repli-seq. Nat Protoc. 13(5):819-839.
Masani S, Han L, Yu K. 2013. Apurinic/apyrimidinic endonuclease 1 is the essential nuclease during immunoglobulin class switch recombination. Mol Cell Biol. 33(7):1468-1473.
Morgan JA, Yin Y, Borowsky AD, Kuo F, Nourmand N, Koontz JI, Reynolds C, Soreng L, Griffin CA, Graeme-Cook F. 1999. Breakpoints of the t (11; 18)(q21; q21) in mucosa-associated lymphoid tissue (MALT) lymphoma lie within or near the previously undescribed gene MALT1 in chromosome 18. Cancer research. 59(24):6205-6213.
Mullighan CG, Collins-Underwood JR, Phillips LAA, Loudin MG, Liu W, Zhang J, Ma J, Coustan-Smith E, Harvey RC, Willman CL. 2009. Rearrangement of CRLF2 in B-progenitor–and Down syndrome–associated acute lymphoblastic leukemia. Nature genetics. 41(11):1243-1246.

Murga Penas EM, Callet-Bauchu E, Ye H, Gazzo S, Berger F, Schilling G, Albert-Konetzny N, Vettorazzi E, Salles G, Wlodarska I et al. 2010. The t(14;18)(q32;q21)/IGH-MALT1 translocation in MALT lymphomas contains templated nucleotide insertions and a major breakpoint region similar to follicular and mantle cell lymphoma. Blood. 115(11):2214-2219.
Nourse J, Mellentin JD, Galili N, Wilkinson J, Stanbridge E, Smith SD, Cleary ML. 1990. Chromosomal translocation t(1;19) results in synthesis of a homeobox fusion mRNA that codes for a potential chimeric transcription factor. Cell. 60(4):535-545.
Offit K, Lo Coco F, Louie DC, Parsa NZ, Leung D, Portlock C, Ye BH, Lista F, Filippa DA, Rosenbaum A et al. 1994. Rearrangement of the bcl-6 gene as a prognostic marker in diffuse large-cell lymphoma. N Engl J Med. 331(2):74-80.
Ohno H. 2004. Pathogenetic role of BCL6 translocation in B-cell non-Hodgkin's lymphoma. Histol Histopathol. 19(2):637-650.
Ohno H. 2011. BCL6 Translocations in B-Cell Tumors. In: Schwab M, editor. Encyclopedia of Cancer. Berlin, Heidelberg: Springer Berlin Heidelberg; p. 364-368.
Pannunzio NR, Lieber MR. 2018. Concept of DNA lesion longevity and chromosomal translocations. Trends in biochemical sciences. 43(7):490-498.
Paulsson K, Jonson T, Ora I, Olofsson T, Panagopoulos I, 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(2):196-201.
Pham P, Bransteitter R, Petruska J, Goodman MF. 2003. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature. 424(6944):103-107.
Pham P, Malik S, Mak C, Calabrese PC, Roeder RG, Goodman MF. 2019. AID–RNA polymerase II transcription-dependent deamination of IgV DNA. Nucleic Acids Research. 47(20):10815-10829.
Pi WC, Wang J, Shimada M, Lin JW, Geng H, Lee YL, Lu R, Li D, Wang GG, Roeder RG et al. 2020. E2A-PBX1 functions as a coactivator for RUNX1 in acute lymphoblastic leukemia. Blood. 136(1):11-23.
Picard F, Cadoret J-C, Audit B, Arneodo A, Alberti A, Battail C, Duret L, Prioleau M-N. 2014. The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells. PLOS Genetics. 10(5):e1004282.
Raffeld M, Jaffe ES. 1991. bcl-1, t (11; 14), and mantle cell-derived lymphomas.
Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, Schreiber SL, Rando OJ, Madhani HD. 2005. Histone variant H2A. Z marks the 5′ ends of both active and inactive genes in euchromatin. Cell. 123(2):233-248.
Reed JC. 1994. Bcl-2 and the regulation of programmed cell death. The Journal of cell biology. 124(1):1-6.
Remstein ED, Kurtin PJ, Buno I, Bailey RJ, Proffitt J, Wyatt WA, Hanson CA, Dewald GW. 2000. Diagnostic utility of fluorescence in situ hybridization in mantle-cell lymphoma. British journal of haematology. 110(4):856-862.
Resnitzky P, Matutes E, Hedges M, Morilla R, Brito-Babapulle V, Khokhar T, Catovsky D. 1996. The ultrastructure of mantle cell lymphoma and other B-cell disorders with translocation t (11; 14)(q13; q32). British journal of haematology. 94(2):352-361.
Rimokh R, Berger F, Delsol G, Charrin C, Berthéas MF, Ffrench M, Garoscio M, Felman P, Coiffier B, Bryon PA et al. 1993. Rearrangement and overexpression of the BCL-1/PRAD-1 gene in intermediate lymphocytic lymphomas and in t(11q13)-bearing leukemias. Blood. 81(11):3063-3067.

Rodic N, Zampella JG, Cornish TC, Wheelan SJ, Burns KH. 2013. Translocation junctions in TCF3-PBX1 acute lymphoblastic leukemia/lymphoma cluster near transposable elements. Mob DNA. 4(1):22.
Ruiz JF, Gómez-González B, Aguilera A. 2011. AID induces double-strand breaks at immunoglobulin switch regions and c-MYC causing chromosomal translocations in yeast THO mutants. PLoS Genet. 7(2):e1002009.
Russell LJ, Capasso M, Vater I, Akasaka T, Bernard OA, Calasanz MJ, Chandrasekaran T, Chapiro E, Gesk S, Griffiths M. 2009. Deregulated expression of cytokine receptor gene, CRLF2, is involved in lymphoid transformation in B-cell precursor acute lymphoblastic leukemia. Blood, The Journal of the American Society of Hematology. 114(13):2688-2698.
Schmutte C, Yang AS, Beart RW, 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 research. 55(17):3742-3746.
Schones DE, Cui K, Cuddapah S, Roh T-Y, Barski A, Wang Z, Wei G, Zhao K. 2008. Dynamic Regulation of Nucleosome Positioning in the Human Genome. Cell. 132(5):887-898.
Schuettengruber B, Chourrout D, Vervoort M, Leblanc B, Cavalli G. 2007. Genome Regulation by Polycomb and Trithorax Proteins. Cell. 128(4):735-745.
Schwalb B, Michel M, Zacher B, Frühauf K, Demel C, Tresch A, Gagneur J, Cramer P. 2016. TT-seq maps the human transient transcriptome. Science. 352(6290):1225-1228.
Shibata E, Kiran M, Shibata Y, Singh S, Kiran S, Dutta A. 2016. Two subunits of human ORC are dispensable for DNA replication and proliferation. Elife. 5.
Shimazaki C, Goto H, Araki S, Tatsumi T, Takahashi R, Hirai H, Kikuta T, Yamagata N, Ashihara E, Inaba T et al. 1997. Overexpression of PRAD1/cyclin D1 in plasma cell leukemia with t(11;14)(q13;q32). Int J Hematol. 66(1):111-115.
Strobl LJ, Kohlhuber F, Mautner J, Polack A, Eick D. 1993. Absence of a paused transcription complex from the c-myc P2 promoter of the translocation chromosome in Burkitt's lymphoma cells: implication for the c-myc P1/P2 promoter shift. Oncogene. 8(6):1437-1447.
Suganuma T, Workman JL. 2011. Signals and combinatorial functions of histone modifications. Annual review of biochemistry. 80:473-499.
Troppan K, Wenzl K, Neumeister P, Deutsch A. 2015. Molecular Pathogenesis of MALT Lymphoma. Gastroenterol Res Pract. 2015:102656-102656.
Tsai AG, Engelhart AE, Ma'mon MH, Houston SI, Hud NV, Haworth IS, Lieber MR. 2009. Conformational variants of duplex DNA correlated with cytosine-rich chromosomal fragile sites. J Biol Chem. 284(11):7157-7164.
Tsai AG, Lu H, Raghavan SC, Muschen M, Hsieh CL, Lieber MR. 2008. Human chromosomal translocations at CpG sites and a theoretical basis for their lineage and stage specificity. Cell. 135(6):1130-1142.
Tsai AG, Yoda A, Weinstock DM, Lieber MR. 2010. 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(11):1993-1994.
Van Attikum H, Gasser SM. 2009. Crosstalk between histone modifications during the DNA damage response. Trends in cell biology. 19(5):207-217.
Wagner SD, Ahearne M, Ferrigno PK. 2011. The role of BCL6 in lymphomas and routes to therapy. British Journal of Haematology. 152(1):3-12.

Walsh CP, Xu GL. 2006. Cytosine methylation and DNA repair. DNA Methylation: Basic Mechanisms. Springer; p. 283-315.
Wang Q, Kieffer-Kwon KR, Oliveira TY, Mayer CT, Yao K, Pai J, Cao Z, Dose M, Casellas R, Jankovic M et al. 2017. The cell cycle restricts activation-induced cytidine deaminase activity to early G1. J Exp Med. 214(1):49-58.
Watts JA, Burdick J, Daigneault J, Zhu Z, Grunseich C, Bruzel A, Cheung VG. 2019. cis elements that mediate RNA polymerase II pausing regulate human gene expression. The American Journal of Human Genetics. 105(4):677-688.
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(4):1629-1636.
Wiemels JL, Leonard BC, Wang Y, Segal MR, Hunger SP, Smith MT, Crouse V, Ma X, Buffler PA, 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 U S A. 99(23):15101-15106.
Wu C. 1980. The 5′ ends of Drosophila heat shock genes in chromatin are hypersensitive to DNase I. Nature. 286(5776):854-860.
Yang Y, McBride KM, Hensley S, Lu Y, Chedin F, Bedford MT. 2014. Arginine methylation facilitates the recruitment of TOP3B to chromatin to prevent R loop accumulation. Molecular cell. 53(3):484-497.
Ye H, Gong L, Liu H, Hamoudi RA, Shirali S, Ho L, Chott A, Streubel B, Siebert R, Gesk S. 2005. MALT lymphoma with t (14; 18)(q32; q21)/IGH-MALT1 is characterized by strong cytoplasmic MALT1 and BCL10 expression. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland. 205(3):293-301.
Yoda A, Yoda Y, Chiaretti S, Bar-Natan M, Mani K, Rodig SJ, West N, Xiao Y, Brown JR, Mitsiades C. 2010. Functional screening identifies CRLF2 in precursor B-cell acute lymphoblastic leukemia. Proceedings of the National Academy of Sciences. 107(1):252-257.
Yu K, Roy D, Bayramyan M, Haworth IS, Lieber MR. 2005. Fine-structure analysis of activation-induced deaminase accessibility to class switch region R-loops. Mol Cell Biol. 25(5):1730-1736.
Zentner GE, Henikoff S. 2013. Regulation of nucleosome dynamics by histone modifications. Nature structural & molecular biology. 20(3):259.
Zhang ZZ, Pannunzio NR, Han L, Hsieh CL, Yu K, Lieber MR. 2014. The strength of an Ig switch region is determined by its ability to drive R loop formation and its number of WGCW sites. Cell Rep. 8(2):557-569.
Zhang ZZ, Pannunzio NR, Hsieh CL, Yu K, Lieber MR. 2014. The role of G-density in switch region repeats for immunoglobulin class switch recombination. Nucleic Acids Res. 42(21):13186-13193.

Figure S1. Mechanism of V(D)J recombination

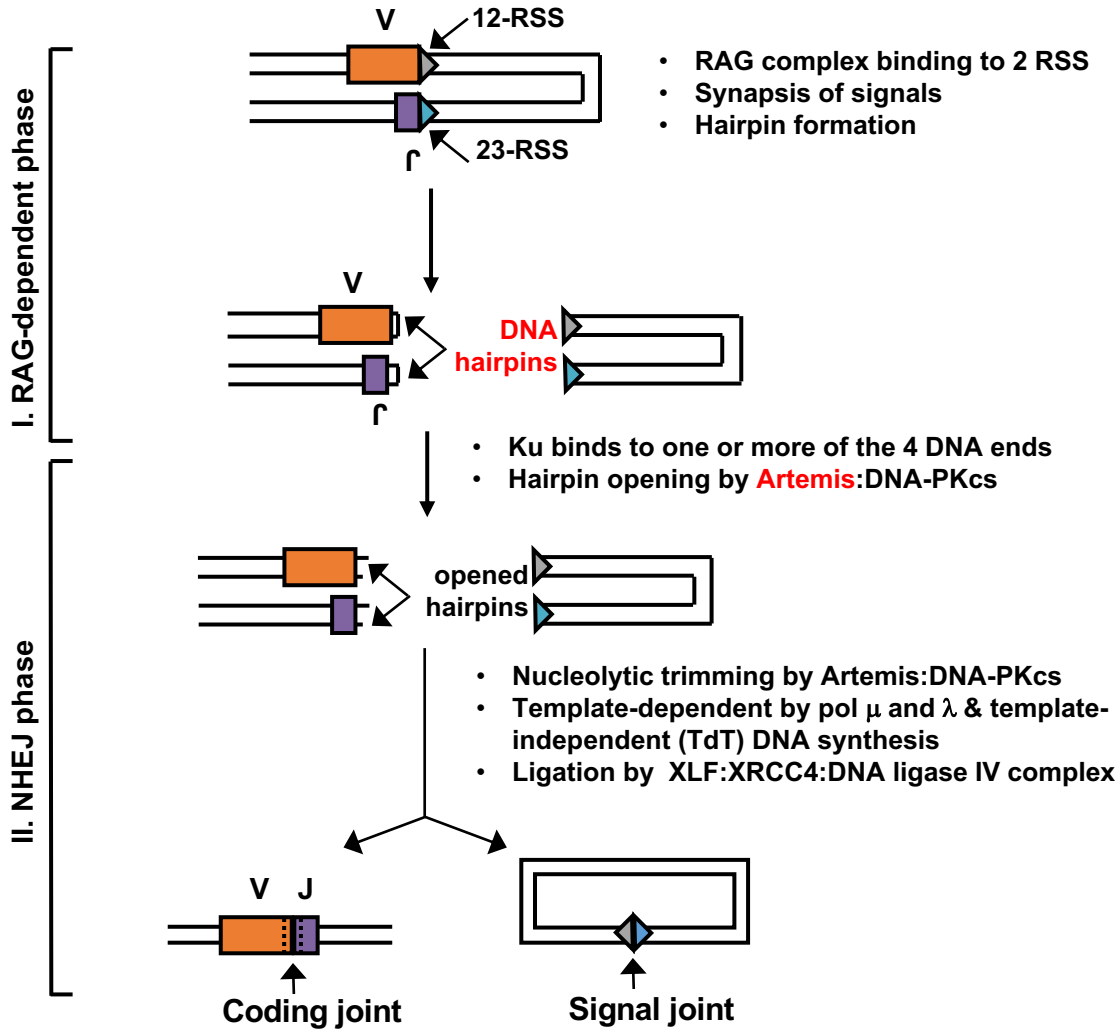


Figure S2. Mechanism of mammalian Ig heavy locus class switch recombination

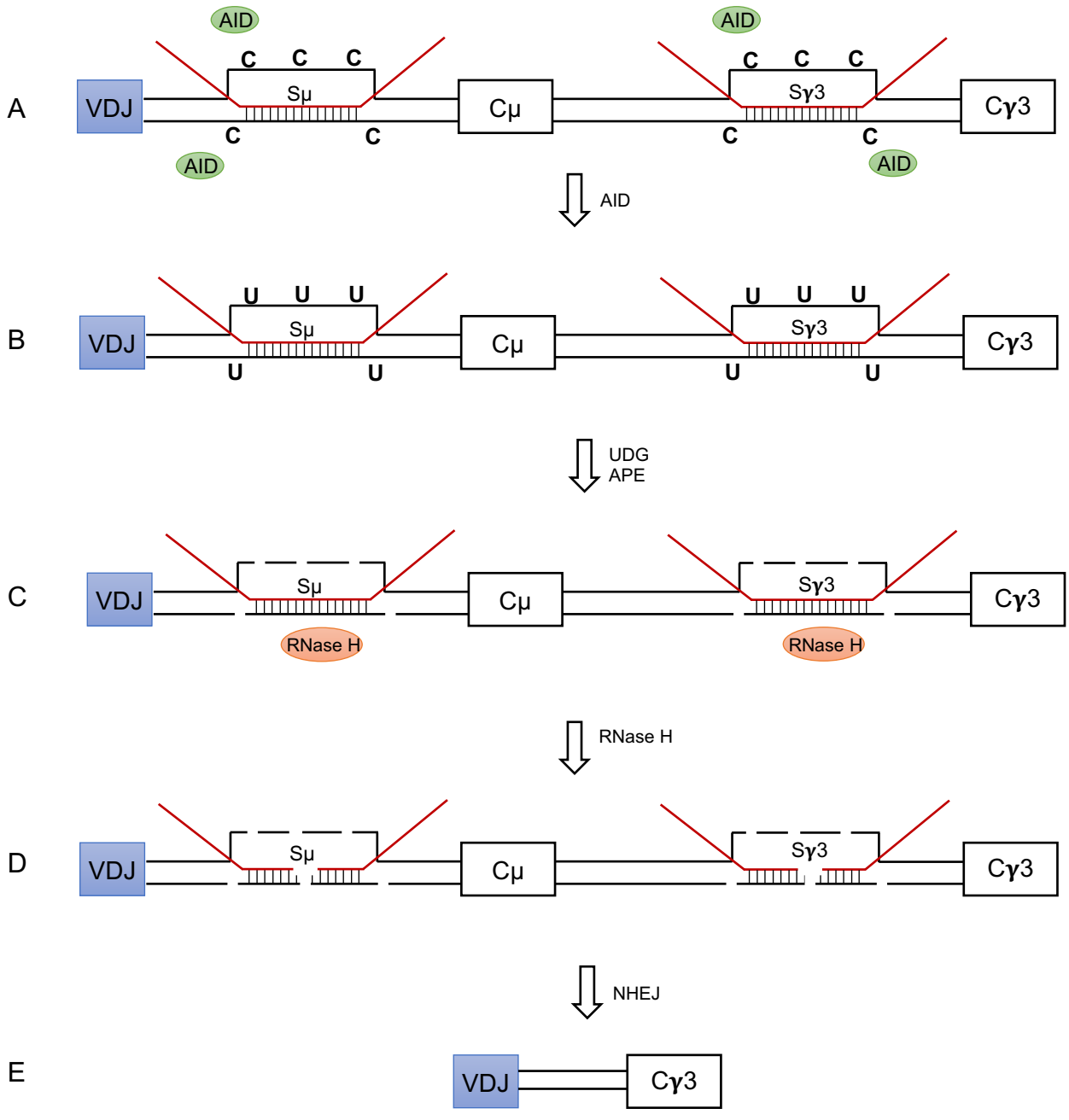


Figure S3. Mechanistic aspects of Ig somatic hypermutation

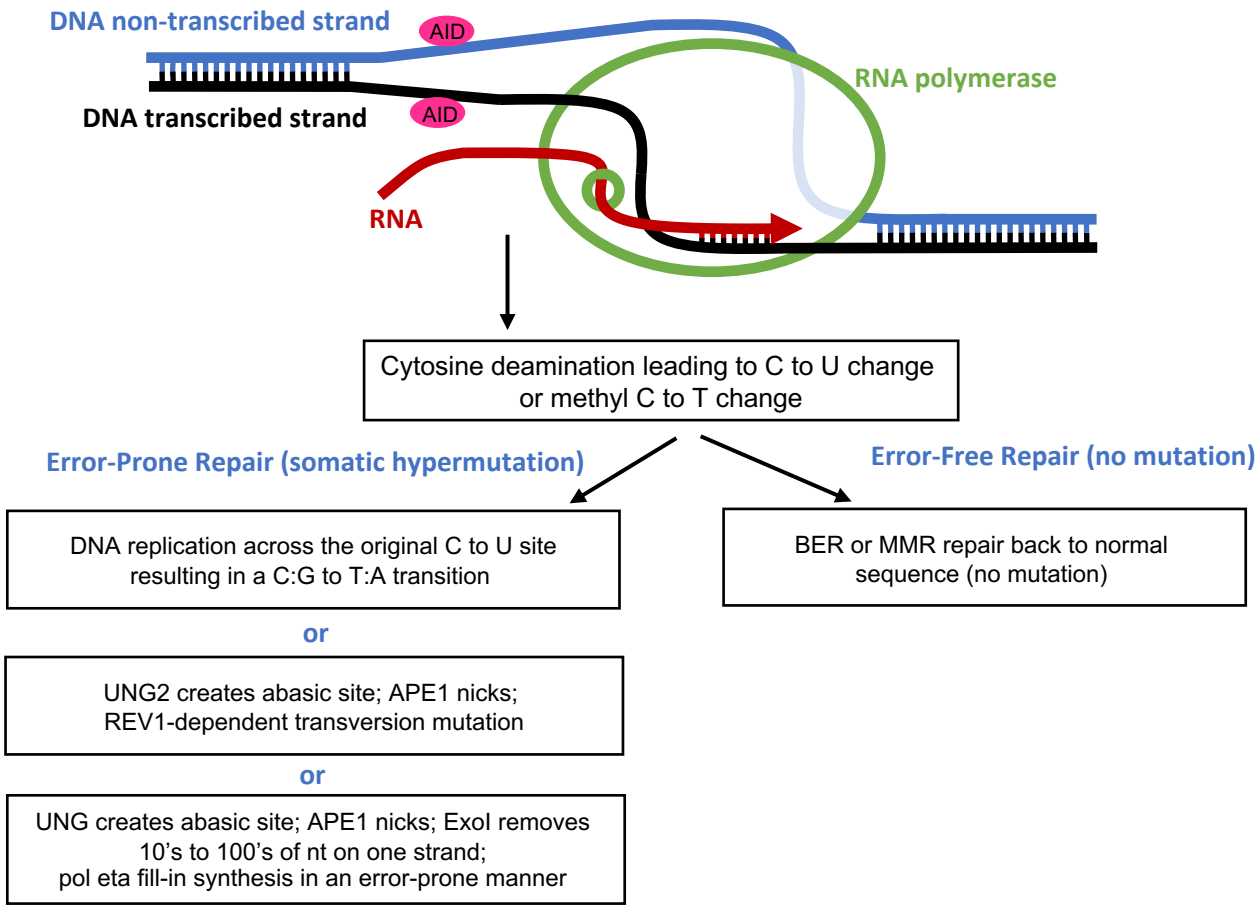


Figure S4. Illustrations of chromosomal translocations

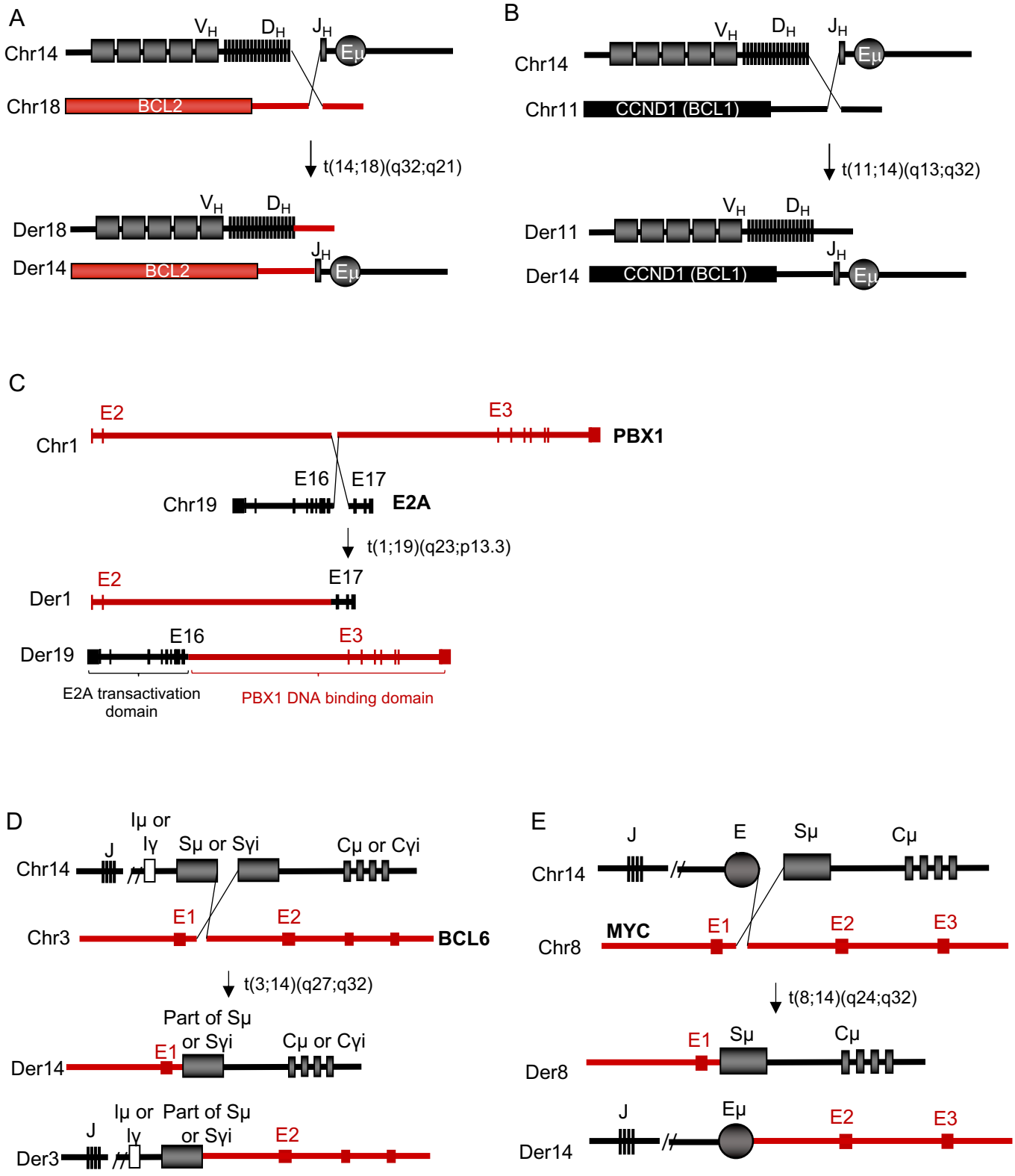
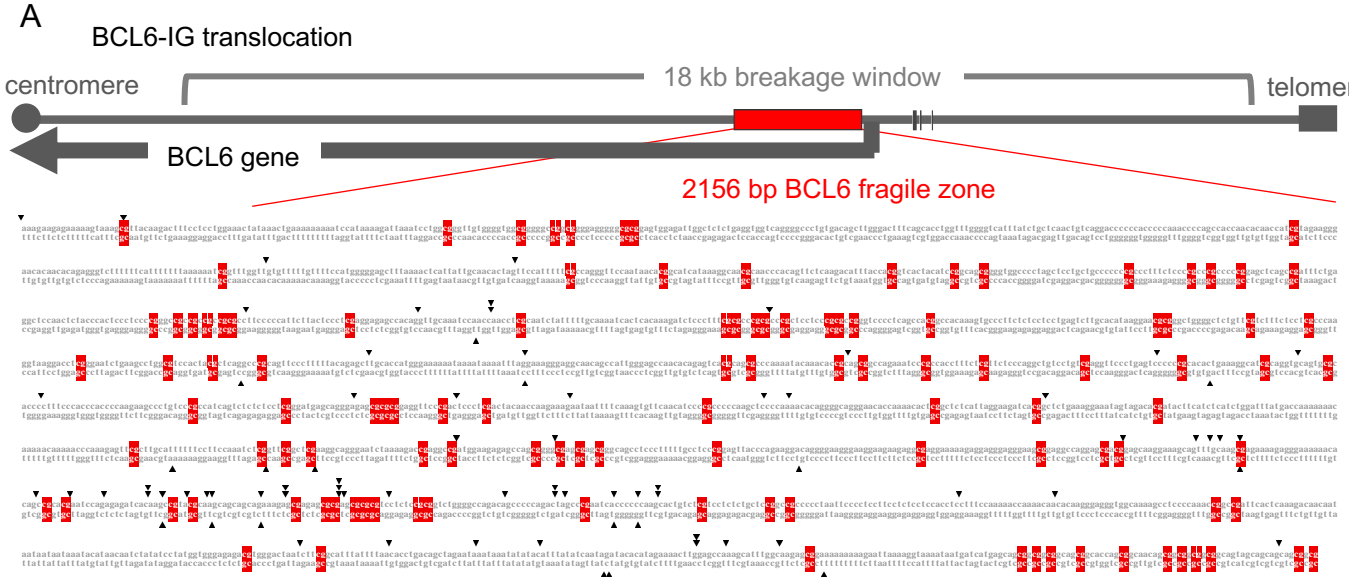
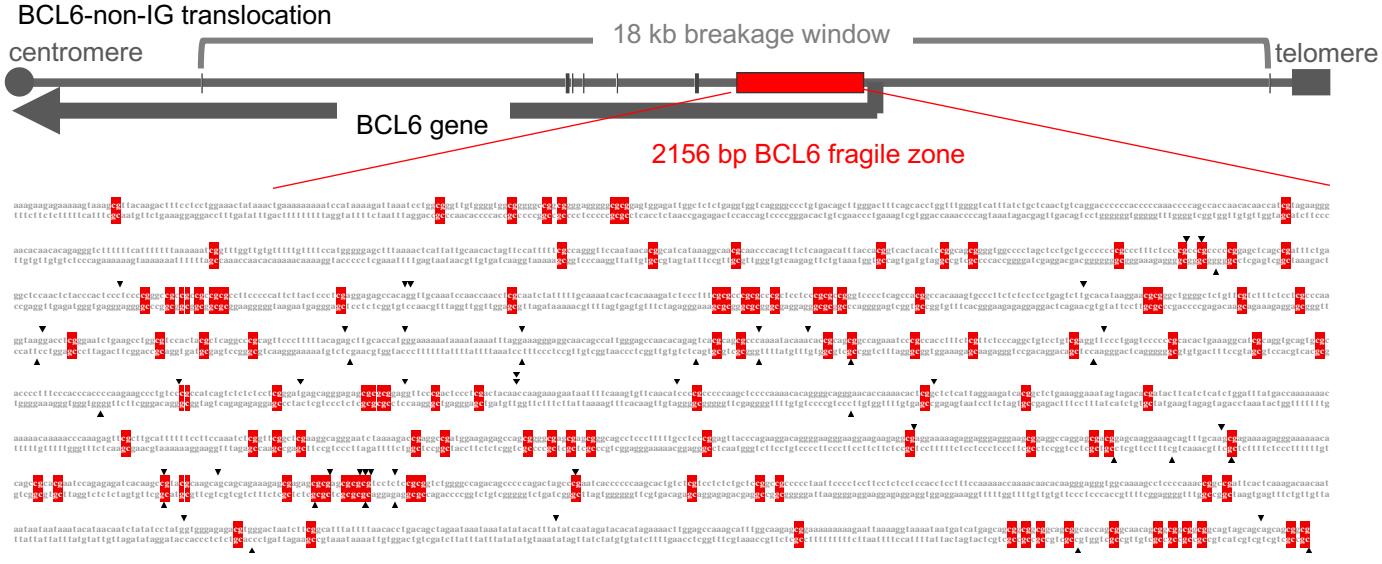


Figure S5. Breakpoint distribution on BCL6 and MYC genes

A





B

Figure S6. Three pathways that lead to DSBs mediated by AID.

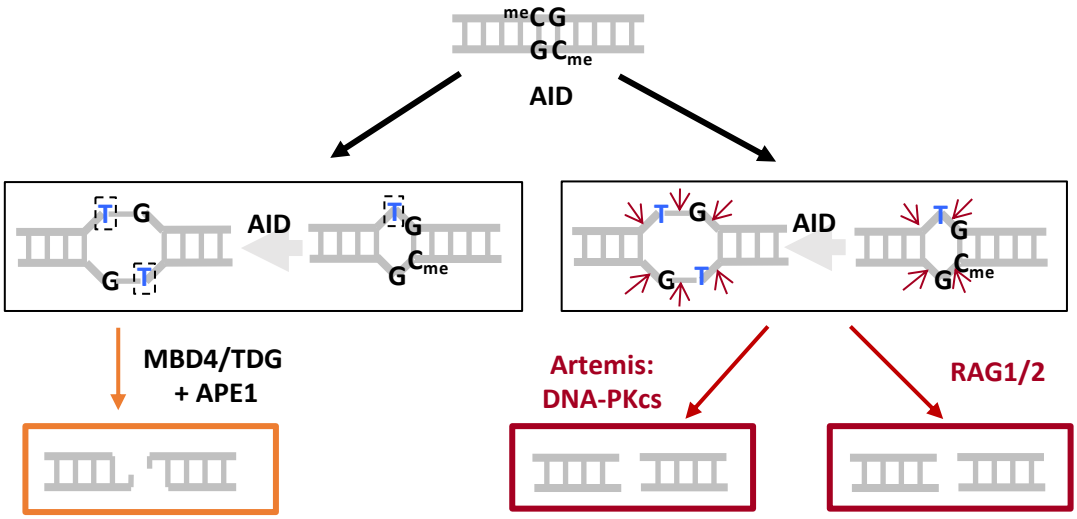


Figure S7. Breakpoints distribution on PBX1 and HLF

A

B

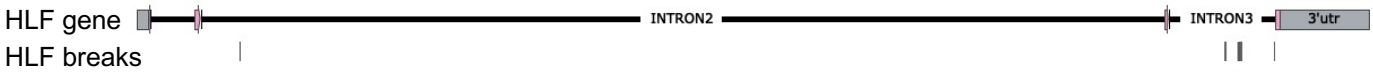
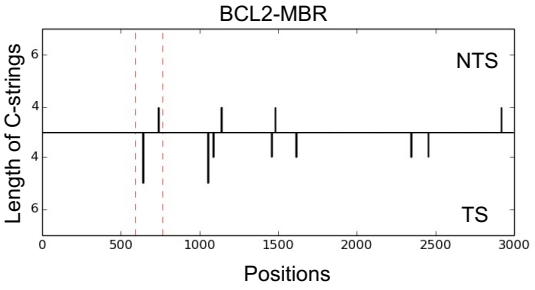
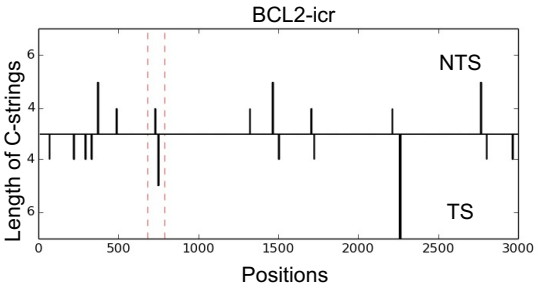
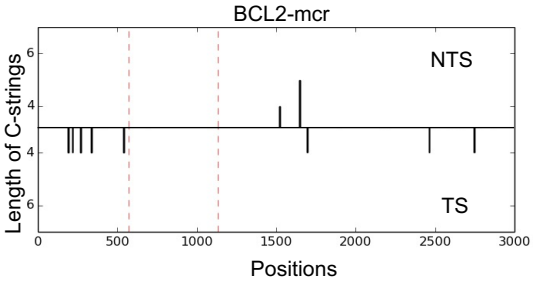


Figure S8. C-string distribution around the fragile zones involved in B cell translocations

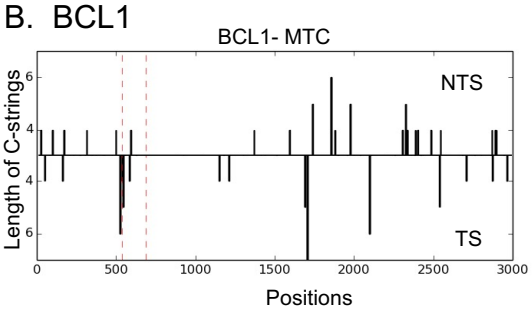
A. BCL2



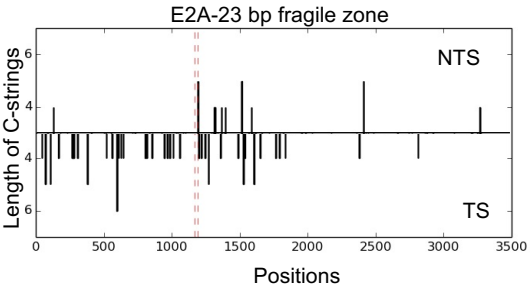




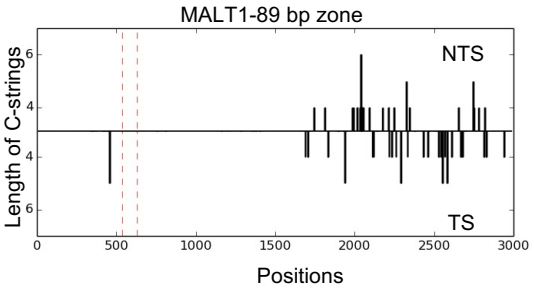
B. BCL1



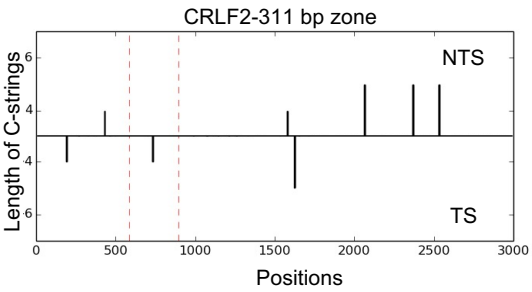
C. E2A



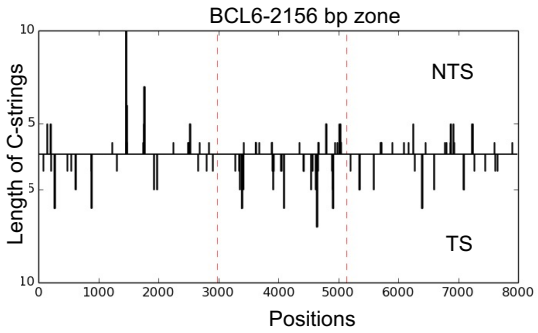
D. MALT1



E. CRLF2



F. BCL6



G. MYC

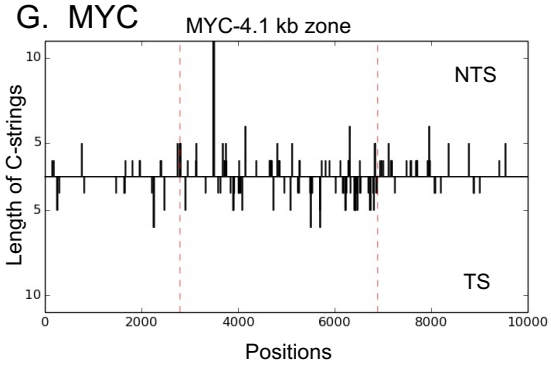


Figure S9. Regional features around the fragile zones of early and mature B cells

A. BCL2

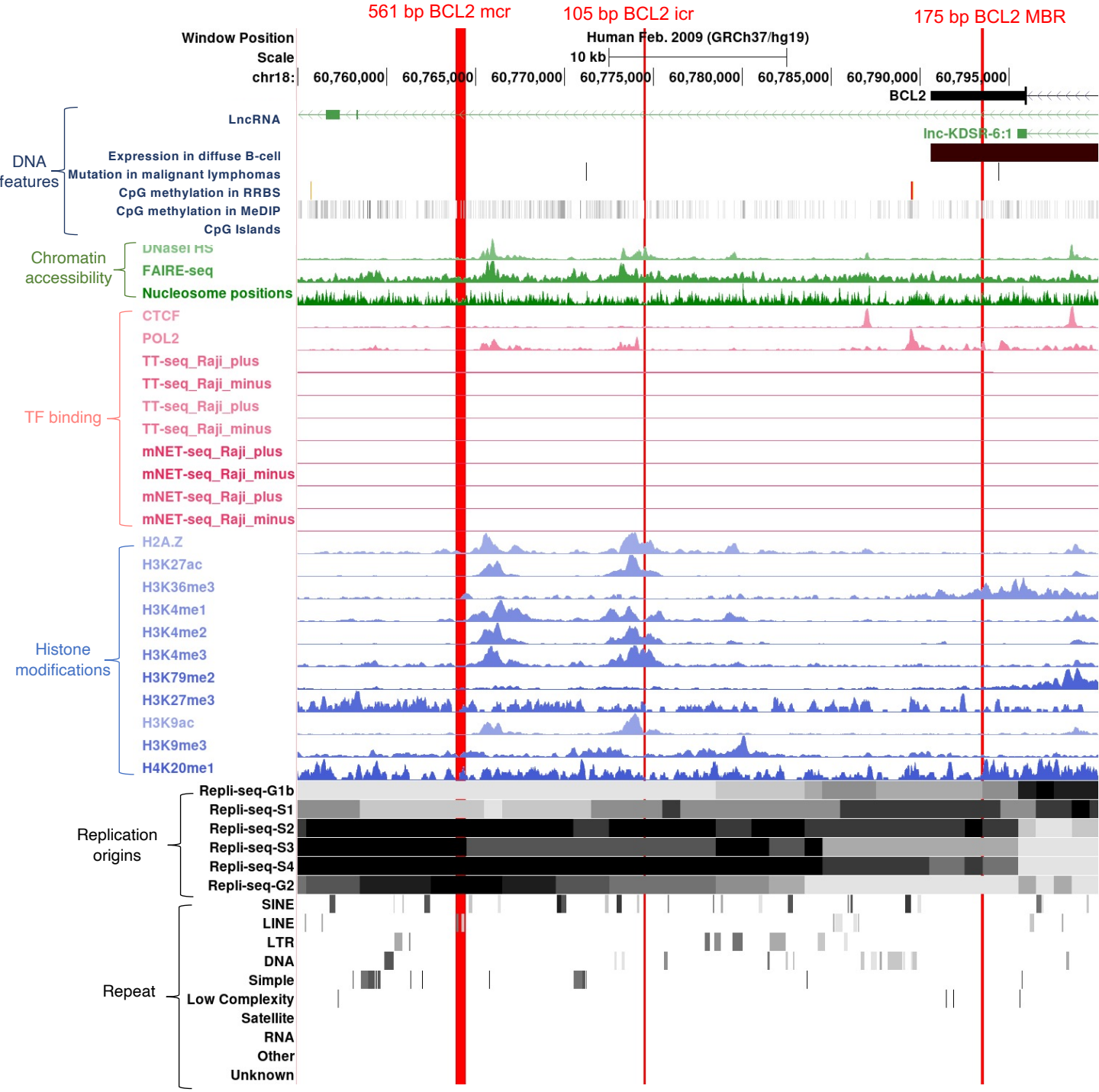


Figure S9 (continued)

B. BCL1

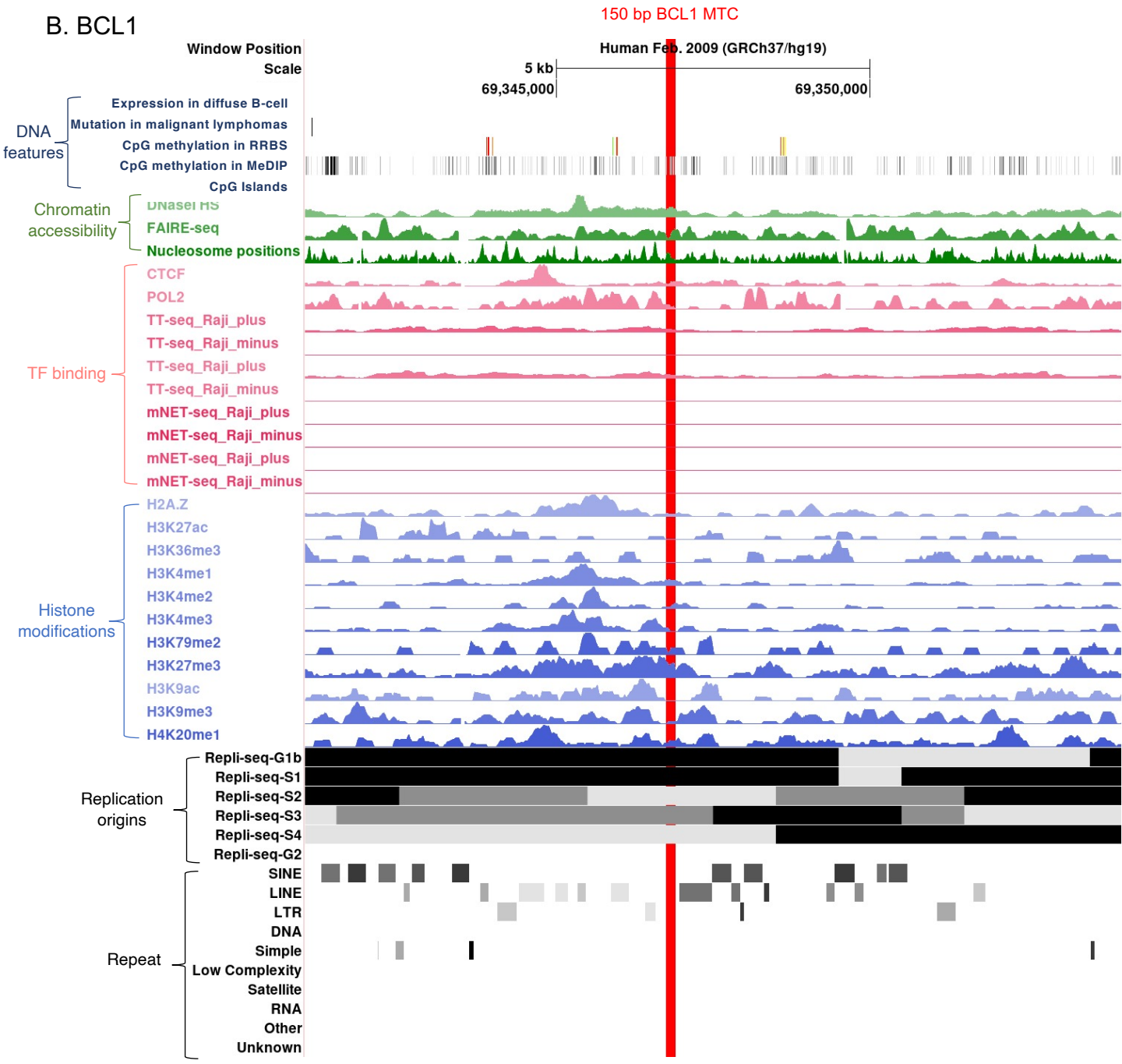


Figure S9 (continued)

C. E2A

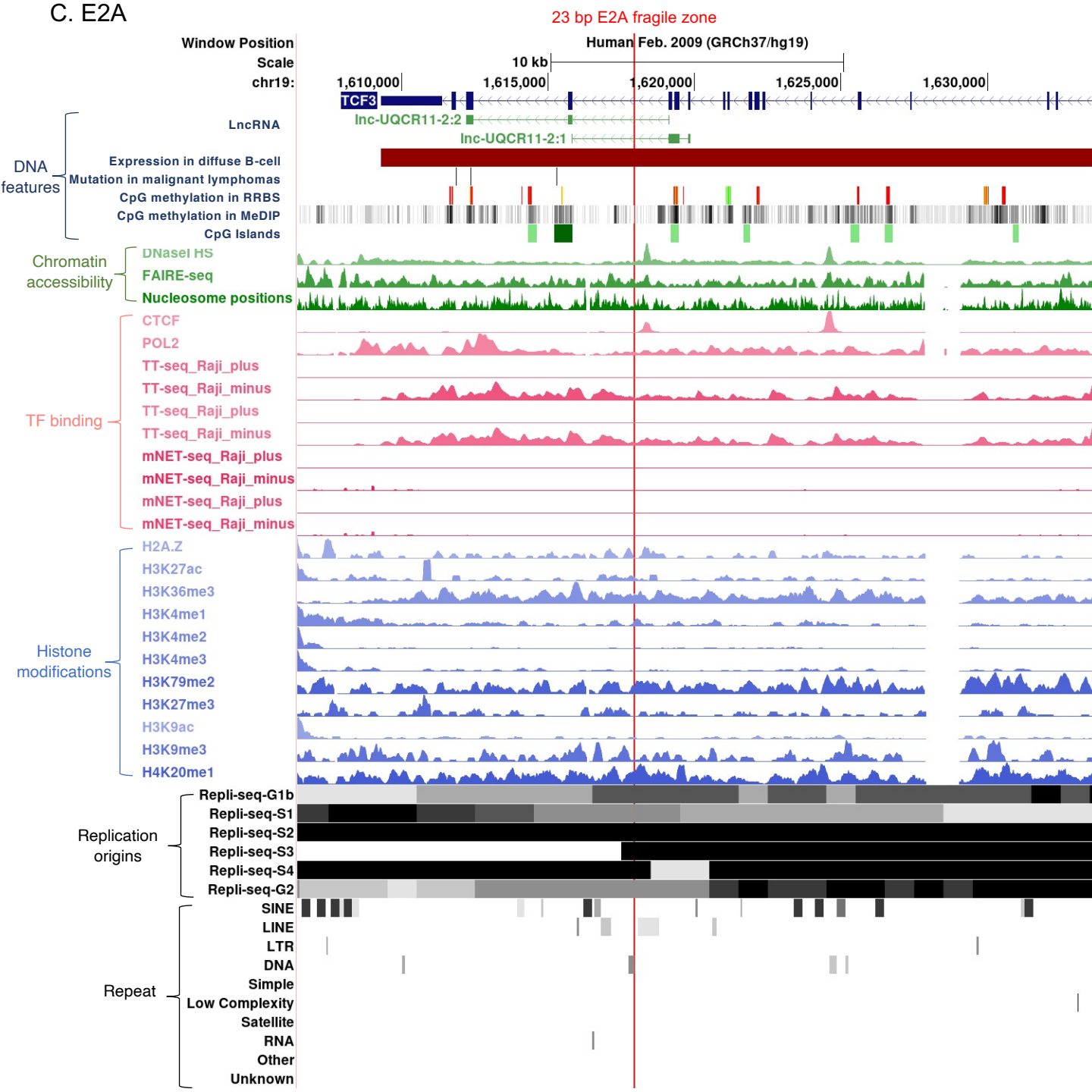


Figure S9 (continued)

D. MALT1

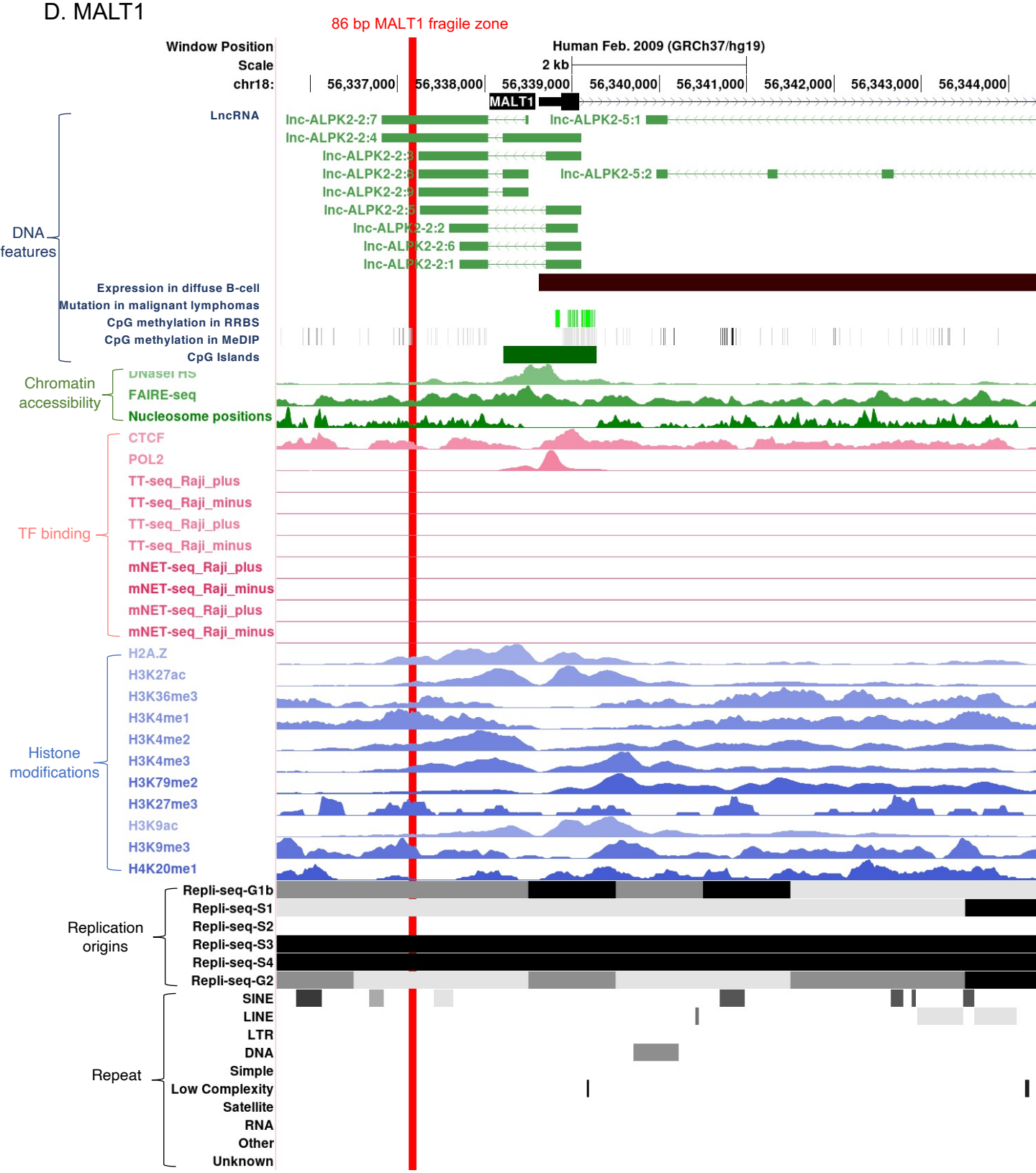


Figure S9 (continued)

E. CRLF2

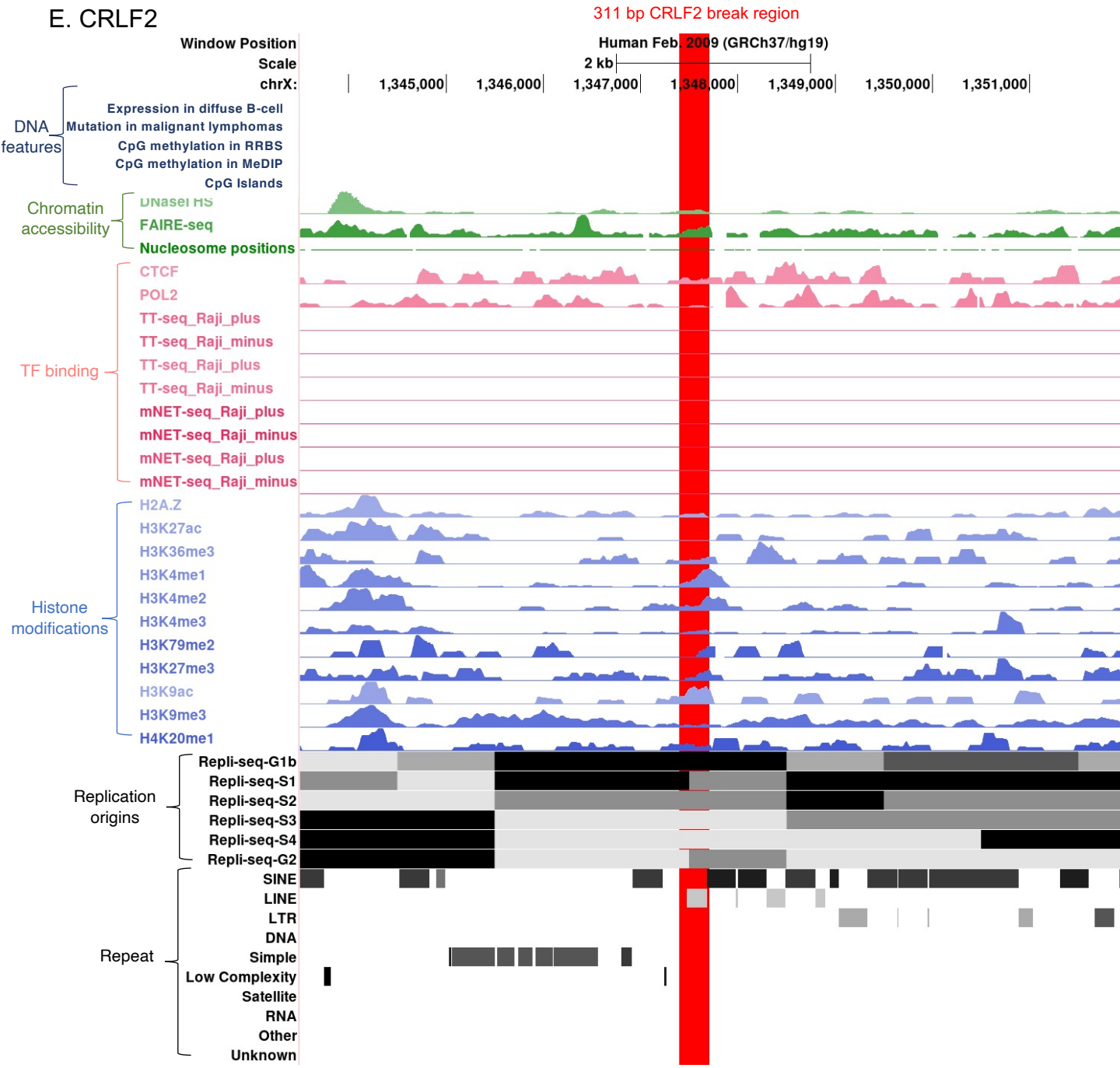


Figure S9 (continued)

F. BCL6

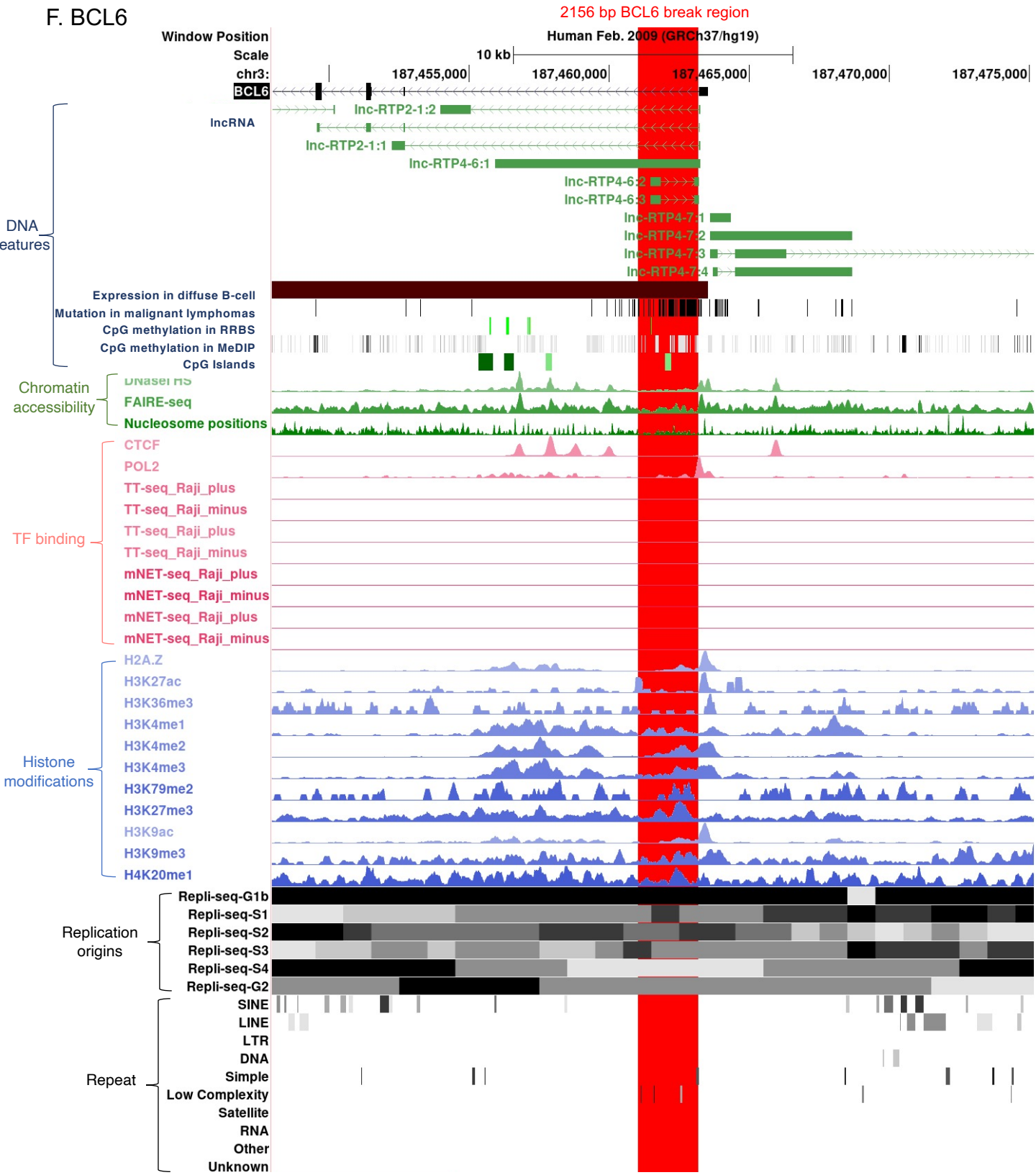


Figure S9 (continued)

G. MYC

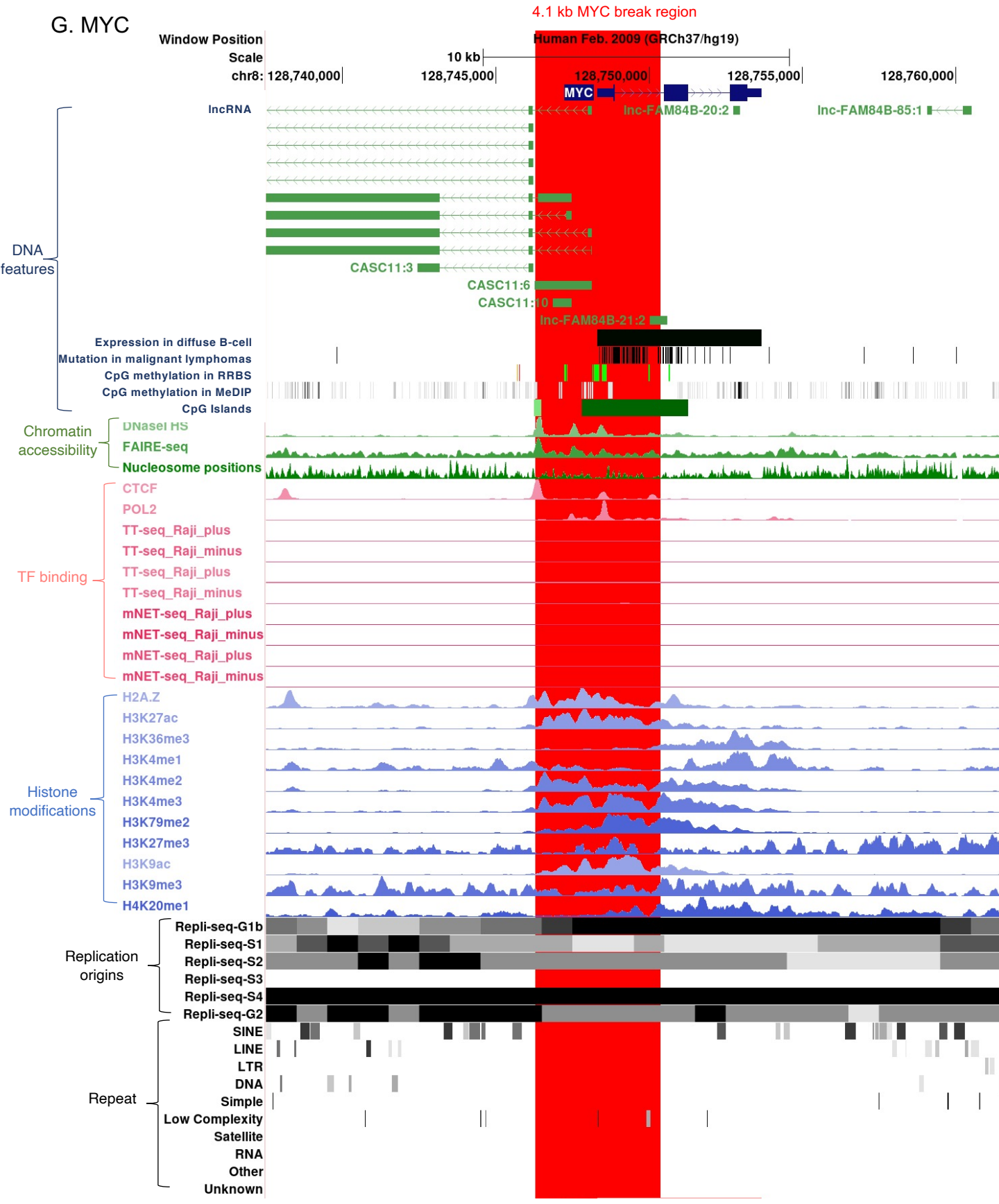


Figure S10. Distribution of DNA repeats near the fragile zones of early B cells

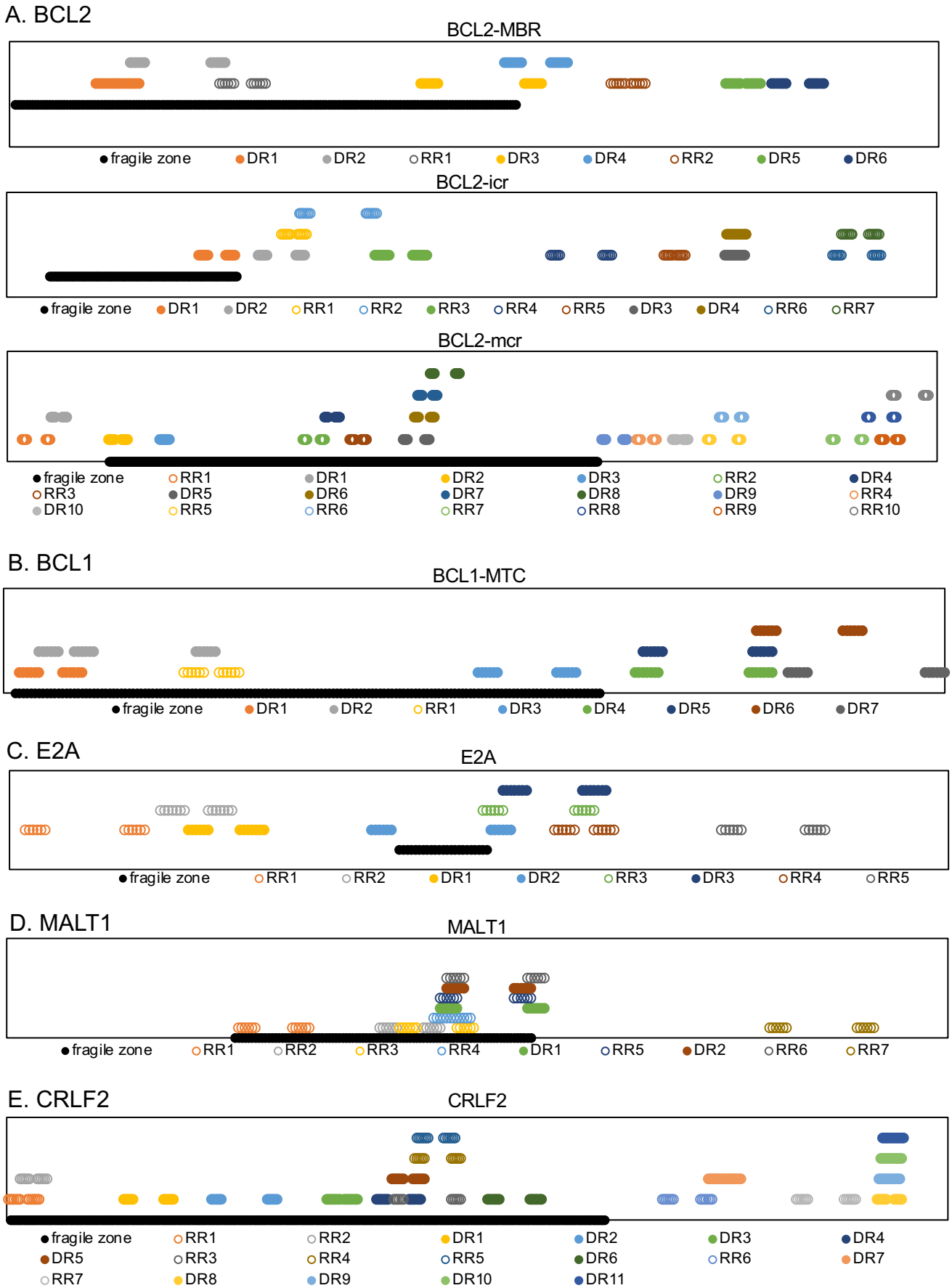
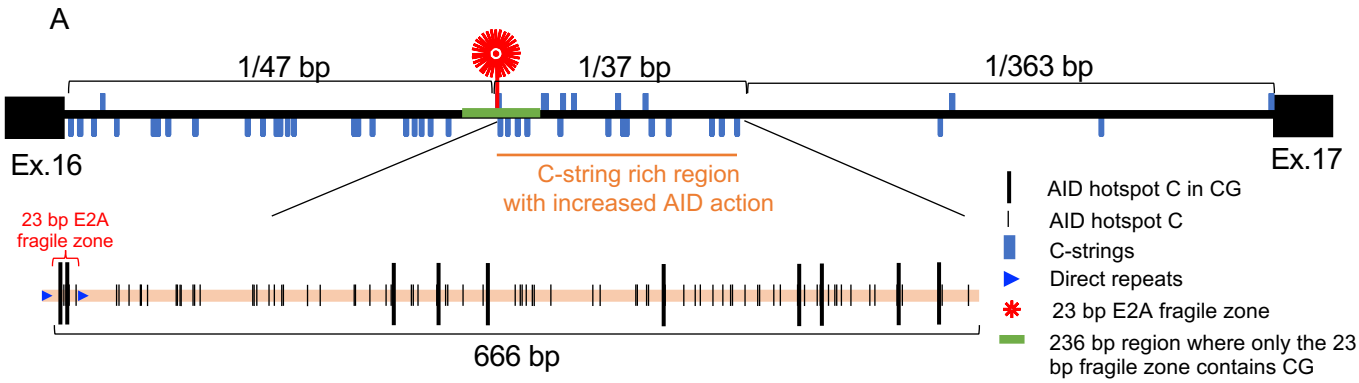


Figure S11. Key factors for DNA breakage in E2A fragile zone



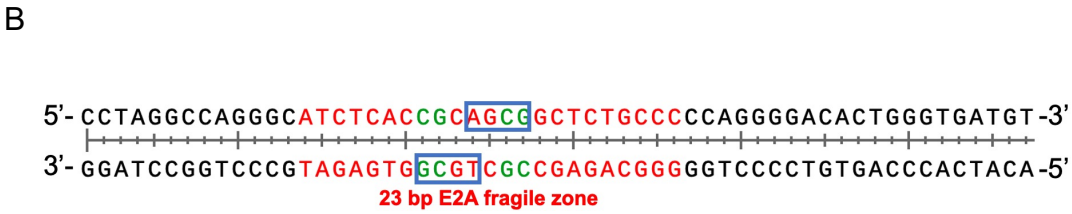


Table S1. Distance of nearby CpG sites nearest to each fragile zone

upstream CG distance in bp	fragile region	downstream CG distance in bp	regions defined by nearest CpG
1	BCL2 175 bp MBR	147	323 bp
23	BCL2 105 bp icr	390	518 bp
116	BCL2 561 bp mcr	390	1067 bp
141	BCL1 150 bp MTC	4	295 bp
97	23 bp E2A fragile zone	116	236 bp
67	MALT1 86 bp fragile zone	124	280 bp
1	CRLF2 311 bp fragile zone	179	491 bp
120	BCL6 2156 bp fragile zone	28	2304 bp
13	MYC 4.1 kb fragile zone	8	4120 bp