Cell

Identification of Early Replicating Fragile Sites that Contribute to Genome Instability

Jacqueline H. Barlow,[1,9] Robert B. Faryabi,[1,9] Elsa Callén,[1] Nancy Wong,[1] Amy Malhowski,[1] Hua Tang Chen,[1] Gustavo Gutierrez-Cruz,[3] Hong-Wei Sun,[4] Peter McKinnon,[6] George Wright,[2] Rafael Casellas,[5] Davide F. Robbiani,[7] Louis Staudt,[2] Oscar Fernandez-Capetillo,[8] and André Nussenzweig[1,*]

[1]Laboratory of Genome Integrity
[2]Metabolism Branch Center for Cancer Research
National Cancer Institute, NIH, Bethesda, Maryland 20892, USA
[3]Laboratory of Muscle Stem Cells and Gene Regulation
[4]Biodata Mining and Discovery Section, Office of Science and Technology
[5]Laboratory of Immunogenetics
National Institute of Arthritis and Musculoskeletal and Skin Diseases, NIH, Bethesda, MD 20892, USA
[6]Department of Genetics, St. Jude Children's Research Hospital, Memphis, TN 38105, USA
[7]Laboratory of Molecular Immunology, The Rockefeller University, New York, NY 10065, USA
[8]Genomic Instability Group, Spanish National Cancer Research Centre (CNIO), E-28029 Madrid, Spain
[9]These authors contributed equally to this work
*Correspondence: andre_nussenzweig@nih.gov
http://dx.doi.org/10.1016/j.cell.2013.01.006

SUMMARY

DNA double-strand breaks (DSBs) in B lymphocytes arise stochastically during replication or as a result of targeted DNA damage by activation-induced cytidine deaminase (AID). Here we identify recurrent, early replicating, and AID-independent DNA lesions, termed early replication fragile sites (ERFSs), by genome-wide localization of DNA repair proteins in B cells subjected to replication stress. ERFSs colocalize with highly expressed gene clusters and are enriched for repetitive elements and CpG dinucleotides. Although distinct from late-replicating common fragile sites (CFS), the stability of ERFSs and CFSs is similarly dependent on the replication-stress response kinase ATR. ERFSs break spontaneously during replication, but their fragility is increased by hydroxyurea, ATR inhibition, or deregulated c-Myc expression. Moreover, greater than 50% of recurrent amplifications/deletions in human diffuse large B cell lymphoma map to ERFSs. In summary, we have identified a source of spontaneous DNA lesions that drives instability at preferred genomic sites.

INTRODUCTION

Double-strand breaks (DSBs) arise spontaneously during DNA replication, as a result of oncogenic stress, and as a part of the gene diversification programs in lymphocytes (Bartek et al., 2007; Callén et al., 2007; Gostissa et al., 2011; Halazonetis et al., 2008). When B lymphocytes are activated, they undergo rapid proliferation and simultaneously initiate two-genome remodeling reactions, termed somatic hypermutation (SHM) and class switch recombination (CSR). The coupling of rapid cycling and programmed DNA damage poses the B cell genome at high risk for destabilization.

SHM introduces point mutations in the variable region of immunoglobulin (Ig) genes, which can increase antibody affinity, whereas CSR is a DNA deletion event that replaces one Ig constant region gene for another. Both of these reactions are initiated by the enzyme activation-induced cytidine deaminase (AID), which deaminates cytosine residues in single-stranded DNA exposed during Ig gene transcription (Chaudhuri and Alt, 2004). In addition to Ig genes, AID causes a considerable amount of collateral genomic damage (Chiarle et al., 2011; Kato et al., 2012; Klein et al., 2011; Liu et al., 2008), including oncogenic targets such as c-Myc (Robbiani et al., 2008). Nevertheless, many recurrent mutations in B cell lymphoma are not associated with AID activity, and the mechanisms of rearrangements at these sites remain unclear.

The DNA damage response (DDR) is activated during programmed rearrangements in lymphocytes to ensure faithful DNA repair and prevent chromosomal translocation (Chen et al., 2000; Petersen et al., 2001). The DDR is also triggered by aberrant oncogene expression that induces precocious entry into S phase and perturbs replication fork progression (Bartek et al., 2007; Bester et al., 2011; Halazonetis et al., 2008). Replication fork instability can also be triggered by exogenous agents such as hydroxyurea (HU), which depletes deoxynucleotide pools, or by deficiencies in homologous recombination pathways that are needed to complete DNA replication after fork stalling or collapse (Schlacher et al., 2012).

Oncogenic stress has been shown to preferentially target genomic regions called common fragile sites (CFSs) (Bartek et al., 2007; Halazonetis et al., 2008). Historically, CFSs have been mapped in lymphocytes but are induced in all cell types under conditions that obstruct replication, such as treatment with low doses of the DNA polymerase inhibitor aphidicolin. DNA breakage within CFSs spans megabase regions. Nevertheless, CFSs share characteristic features including association with very large genes, enrichment of long stretches of AT dinucleotide-rich repeats, and incomplete DNA replication (Durkin and Glover, 2007).

Replication-stress-induced DNA damage is also observed in yeast. Similar to CFSs, sites located in "replication slow zones" (RSZs) are late replicating and breakage prone (Cha and Kleckner, 2002). In addition to late replicating areas, irreversible replication fork collapse in response to acute doses of hydroxyurea has been observed preferentially around a subset of early firing replication origins in yeast (Raveendranathan et al., 2006), which do not overlap with RSZs (Cha and Kleckner, 2002; Hashash et al., 2011). Although the molecular mechanisms governing replication initiation in yeast and mammalian cells are distinct, we wondered if fragility at early firing origins is also a feature of mammalian cells. Here, we identify highly unstable regions of the B cell genome designated as "early replicating fragile sites" (ERFSs). We propose that ERFSs are a new class of fragile sites in mammalian cells that contribute to recurrent rearrangements during lymphomagenesis.

RESULTS

Genome-wide Mapping of Replication-Induced DNA Damage

Single-strand DNA (ssDNA) mapping has been used to localize origins of replication in yeast (Feng et al., 2006). To identify potential sites of fork collapse, we first profiled the location and extent of ssDNA genome-wide using chromatin immunoprecipitation (ChIP) with an anti-replication protein A (RPA) antibody (Figure 1). RPA associates with ssDNA at stalled forks near early firing origins when fork movement is inhibited by HU (Tanaka and Nasmyth, 1998).

Freshly isolated mouse B cells are arrested in the G_0 phase of the cell cycle (Figure 1A). Upon stimulation with LPS/IL4, cells synchronously enter into the cell cycle so that by 22 hr, approximately 8% of cells have entered S phase, whereas at 28 hr over 30% are in S/G2 phases (Figure 1A). To profile early replication origins, we treated cells at 22 hr with 10 mM HU for 6 hr to fully arrest cells at G_1/S (Figures 1A and 1B). We then performed ChIP-seq of RPA in both untreated and HU-treated cells at 28 hr (Figures 1A and 1B). Two independent experiments showed reproducibility of genome-wide RPA association in HU-treated cells (Figure S1A available online). We generated profiles of RPA in untreated and treated samples, centered on individual RPA-bound sites (Figure S1B), and observed a marked increase in the intensity of RPA in HU-treated B cells relative to untreated cells where 5,939 out of 11,942 genomic regions (49.7%) displayed more than a 4-fold increase in RPA recruitment. In addition to the 53% overlap of RPA-associated regions between HU-untreated versus -treated cells, we also observed that 1,441 regions were present only in HU-treated samples (Figure S1B). These HU-dependent ssDNA regions may correspond to the firing of new replication origins to compensate for inefficient replication.

To confirm that RPA recruitment maps early replication zones, we used the Repli-Seq approach (Hansen et al., 2010) to identify replication origins in B cells during HU arrest. Approximately 12,000 early activating replication origins across the murine B cell genome were identified (Figure S1C). By comparing the distribution of BrdU incorporation relative to the individual RPA-occupied genomic regions, we observed association of BrdU incorporation with nearly 80% of RPA-bound regions (Figure S1C). Moreover, more than 86% of RPA/BrdU enriched genomic sites coincided with previously mapped early replicating regions in the mouse B cell line CH12 (Stamatoyannopoulos et al., 2012) ($p(permutation) < 1 \times 10^{-5}$, Figure S1D). Thus, HU-arrested B cells exhibited an enrichment of RPA at early replicating zones, consistent with an early S phase cell-cycle arrest (Figure 1A).

Early replicating regions are associated with accessible chromatin configuration (MacAlpine et al., 2004). In agreement with this, we found that more than 67% of RPA-bound regions in HU-arrested cells reside within intragenic sequences (Figures S1E and S1I), a frequency significantly higher than expected ($p(permutation) < 1 \times 10^{-5}$). Moreover, RPA preferentially associated with DNaseI hypersensitive sites (DHS) and euchromatic promoters marked by H3K4me3 (Figure S1F). Finally, we measured transcriptional activity in HU-treated B cells directly by genome-wide RNA sequencing. We observed high transcription activity within the RPA-occupied genomic regions as shown by the aggregated pattern of RNA-Seq centered on those regions (Figure S1G). Moreover, 6,100 RPA-bound RefSeq genes exhibited significantly higher average mRNA abundance than those that did not show RPA binding ($p < 1 \times 10^{-16}$, Figure S1H). Thus, HU-induced RPA recruitment in early S phase maps to actively transcribed genes that show the hallmarks of euchromatin.

Replisome stalling in response to HU triggers the activation of the ATR kinase (Ward and Chen, 2001), which protects forks from collapse (Cimprich and Cortez, 2008), and leads to phosphorylation of H2AX (γ-H2AX) (Ward and Chen, 2001), which colocalizes with RPA (Petermann et al., 2010). To examine the relative distribution of γ-H2AX and RPA genome-wide, we carried out ChIP-seq with an antibody that recognizes γ-H2AX (Figure S1A) and examined their profiles with respect to the center of RPA-bound sites. γ-H2AX-associated genomic regions were much broader than RPA, but these regions overlapped with 93% of RPA-bound sites marking ssDNA in HU-treated cells (Figure 1C), consistent with the finding that γ-H2AX marks stalled forks even prior to DSB formation (Petermann et al., 2010). γ-H2AX/RPA enriched loci may therefore correspond to a combination of stalled and broken replisomes.

Cells deficient in homologous recombination (HR) pathway components, such as XRCC2, often accumulate spontaneous chromosome breaks and exhibit hypersensitivity to HU (Sonoda et al., 1998). Consistent with increased spontaneous DNA damage at replication forks, untreated *XRCC2*$^{-/-}$ cells exhibited accumulation of γ-H2AX at similar genomic regions and at almost similar levels observed in HU-treated wild-type (WT)

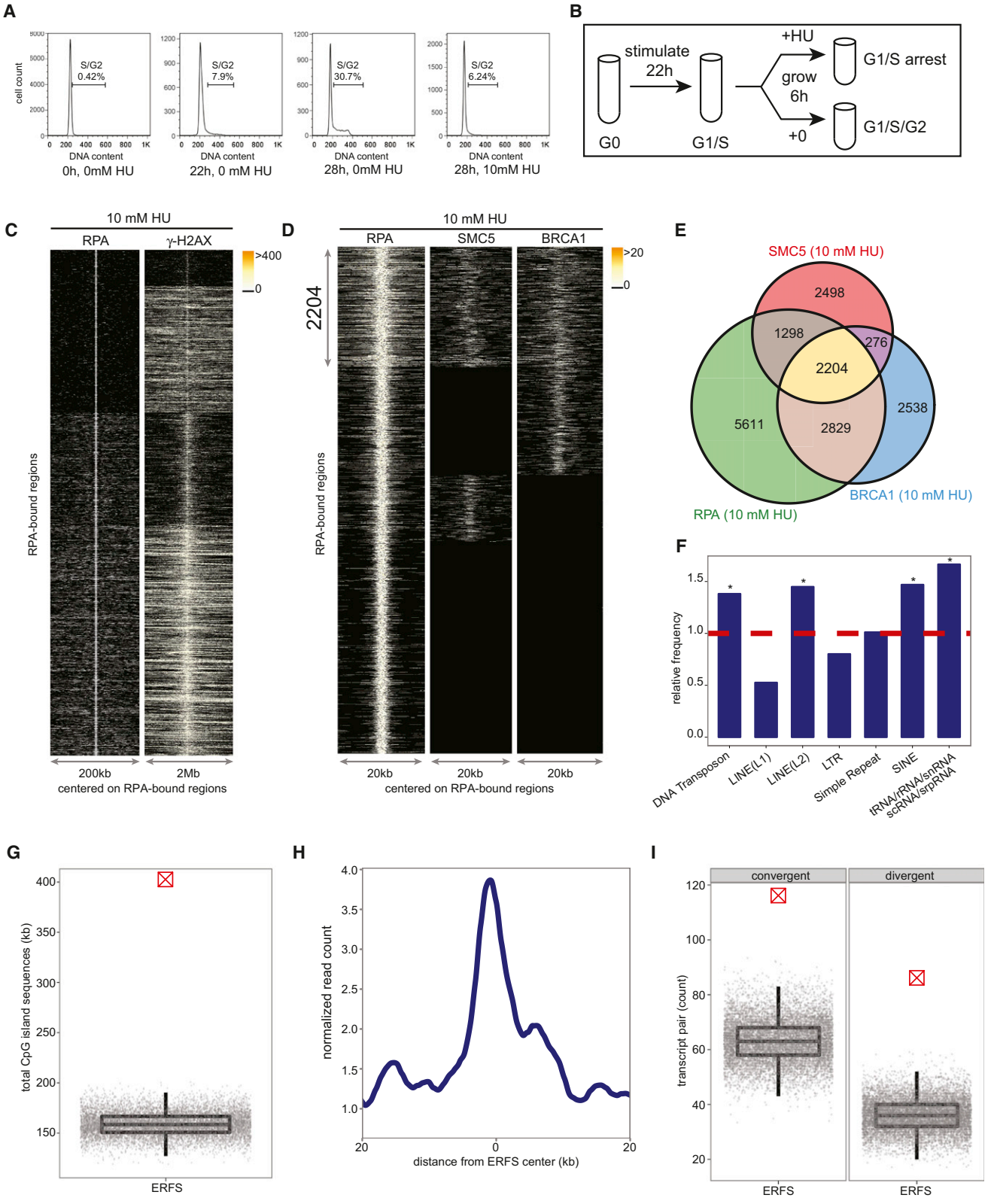


Figure 1. Mapping Replication-Induced DNA Damage in Murine B Lymphocytes

(A) FACS analysis showing DNA content of freshly isolated and ex vivo stimulated splenic murine B lymphocytes in the absence and presence of 10 mM HU.

(B) Experimental plan describing cell synchronization and isolation for samples used in ChIP-seq and RNA-Seq experiments.

(legend continued on next page)

B cells (Figures S2A–S2C). 90% of γ-H2AX-associated genomic regions in untreated *XRCC2*$^{-/-}$ cells correlate with the regions enriched for this protein in HU-treated WT B cells (Figure S2B), and nearly 80% of the regions with enriched γ-H2AX observed in HU-treated WT B cells overlapped with those seen in HU-treated *XRCC2*$^{-/-}$ cells (Figure S2C). These data indicate that XRCC2 deficiency leads to increased endogenous levels of replication stress mostly at the same loci where HU induces replication fork stalling and/or breakage in WT cells.

RPA, BRCA1, and SMC5 Colocalization Marks the Sites of Replication Stress in Early Replicating Zones

Like XRCC2, BRCA1 and members of the structural maintenance of chromosome (SMC) family have been implicated in promoting replication fork restart (Schlacher et al., 2012; Stephan et al., 2011). To determine whether HR proteins bind to a subset of stalled forks marked by RPA and γ-H2AX, we also defined the genome-wide profile of BRCA1 and SMC5. We confirmed BRCA1 and SMC5 ChIP-seq efficacy by observing their association at both Sμ and Sγ1 in *53BP1*$^{-/-}$ cells, where the breaks in IgH persist unrepaired and undergo extensive resection (Figure S3A) (Bothmer et al., 2010; Bunting et al., 2010, 2012; Yamane et al., 2011, 2013).

We then determined the localization of BRCA1 and SMC5 in HU-arrested B cells. Two independent experiments showed reproducibility of genome-wide BRCA1 and SMC5 association (Figures S3B and S3C). To identify the RPA genomic sites co-occupied by the HR proteins BRCA1 and SMC5, we plotted the distribution of their binding with respect to the center of individual RPA-bound regions. Overall, 2,204 regions spanning 10 kbp on average showed RPA/BRCA1/SMC5 triple colocalization (Figures 1D and 1E). We found that RPA was recruited to more than 88% of genomic sites exhibiting BRCA1 and SMC5 association (Figure 1E). Furthermore, genome-wide analysis of RPA/BRCA1/SMC5 profiles in untreated cells revealed more than a 21% increase in the number of genomic regions occupied by these three proteins after HU treatment (Figure S4A). Nevertheless, 48% of RPA/BRCA1/SMC5 triple colocalizations were common between the unperturbed and HU-arrested B cells (Figure S4A). Therefore, we hypothesized that chromatin with concomitant RPA, BRCA1, and SMC5 binding might correspond to regions undergoing replication fork collapse both in response to replication stress and during normal DNA replication. Given that our analysis focused on early replicating sites, which contrasts with late replicating CFSs, we designated these regions as ERFSs.

We then characterized ERFSs to determine whether they share common underlying primary sequence characteristics. Indeed, these loci were enriched at known repetitive elements, including LINE L2, SINE, DNA transposons, and tRNA elements (p(permutation) $< 1 \times 10^{-3}$, Figure 1F), which are known replication fork barriers (Mirkin and Mirkin, 2007). Furthermore, ERFSs showed significantly higher G and C nucleotide content compared to the whole mouse genome, in contrast to CFSs that are enriched in A+T sequences (p(Wilcoxon) $< 1 \times 10^{-16}$, Figure S4B). Twenty-six percent of the ERFSs regions overlapped with CpG islands, which are highly enriched at translocation breakpoints in B cell lymphoma (Tsai et al., 2008). Conversely, CpG islands covered approximately 400,000 nucleotides within these regions (p(permutation) $< 1 \times 10^{-5}$, Figure 1G). As anticipated, ERFSs clustered at early replication origins (Figure S4C), and over 66% of the loci overlapped with intragenic or promoter sequences of RefSeq annotated protein coding genes (p(permutation) $< 1 \times 10^{-3}$, Figures S4D and S4E). Moreover, ERFSs are more transcriptionally active relative to flanking genomic regions shown by relative mRNA enrichment by RNA-Seq (Figure 1H). Indeed, more than 86% of the RefSeq annotated genes with ERFSs are among the highest transcribed genes (p(binomial) $< 1 \times 10^{-16}$, Figure S4F). Finally, ERFSs were significantly enriched in gene pairs that are transcribed in converging or diverging directions (see Experimental Procedures), such as the convergent transcription pair of *IKZF1* and *FIGNL1* shown in Figure 2A. Compared to expected values, ERFSs were at least two times more likely to localize in regions containing gene pairs exhibiting convergent and/or divergent gene pairs (p(permutation) $< 1 \times 10^{-5}$, Figure 1I).

(C) For each RPA-bound site in response to 10 mM HU (y axis), each column depicts the presence of RPA (left) and γ-H2AX (right) within a window centered on the RPA-bound sites. Color map corresponds to binding intensities where "black" represents no binding. K-mean clustering algorithm was used to group the protein-bound sites.

(D) RPA, SMC5, and BRCA1 co-occupy 2,204 genomic regions in response to 10 mM HU. The plot in each column, from left to right, represents the pattern of RPA, SMC5, and BRCA1 genomic occupancy in response to HU centered on RPA-bound sites. K-mean clustering algorithm is used to group the protein-bound sites.

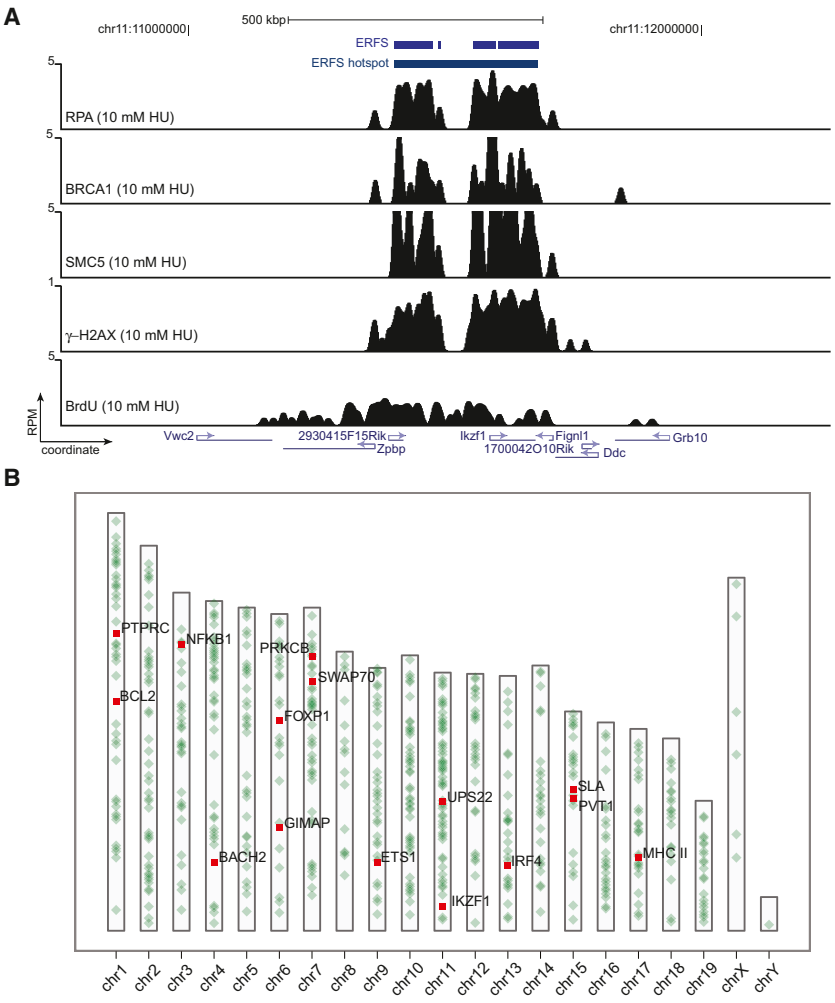
(E) The Venn diagram shows the overlap of sites bound by RPA, SMC5, and BRCA1 in response to 10 mM HU. The total number of bound sites is indicated for each shared and unique area.

(F) Relative frequency of ERFSs in classes of repetitive sequences is shown. Dashed line indicates the expected frequency based on the permutation model (*, enriched repetitive element classes; $p < 1 \times 10^{-3}$).

(G) ERFSs are enriched in CpG islands. Total CpG island sequences in all the 2,204 ERFSs as indicated by the crossed red point is compared to the permutation model as indicated by the gray points. Each gray point corresponds to the total CpG island sequences covered in an iteration of the permutation model. The box plot depicts the quantiles of total CpG sequences based on the permutation model ($p < 1 \times 10^{-5}$).

(H) ERFS genomic regions are transcriptionally active. The line plot represents the average RNA tag count (loess smoothed) in a genomic window around the center of the ERFSs.

(I) ERFSs are enriched in transcriptionally active convergent and divergent gene pairs. Count of divergent/convergent gene pairs coinciding with ERFSs as indicated by the crossed red point is compared to the permutation model as indicated by the gray points. Each gray point corresponds to the total number of divergent/convergent gene pairs observed in an iteration of the permutation model. The box plot depicts the quantiles of the total convergent/divergent transcript pair count based on the permutation model ($p < 1 \times 10^{-5}$). For definition of convergent/divergent gene pairs see Experimental Procedures. See also Figures S1, S3, S4.



C

	ERFS Name	Chromosome	Length (Kb)	Gene Cluster	Transcribed	Convergent and/or Divergent Transcripts	Rearranged in B Cell Cancer	AID Target
1	MHCII	17	333	Yes	Yes	Yes	Yes	Yes
2	GIMAP	6	128	Yes	Yes	Yes	No	Yes
3	SWAP70	7	314.6	Yes	Yes	Yes	No	No
4	BACH2	4	419.4	No	Yes	Yes	Yes	No
5	IKZF1	11	281.4	Yes	Yes	Yes	Yes	Yes
6	FOXP1	6	431.8	No	Yes	No	Yes	No
7	PTPRC	1	1121.6	Yes	Yes	Yes	No	No
8	SLA	15	498.8	Yes	Yes	Yes	No	No
9	PVT1	15	669.8	No	Yes	Yes	Yes	Yes
10	ETS1	9	520	Yes	Yes	Yes	Yes	Yes
11	IRF4	13	283	Yes	Yes	Yes	Yes	Yes
12	BCL2	1	656.8	Yes	Yes	Yes	Yes	No
13	USP22	11	60.2	No	Yes	Yes	No	No
14	NFκB1	3	117	No	Yes	No	No	Yes
15	PRKCB	7	170.6	No	Yes	No	No	No

Figure 2. ERFS "Hot Spots" Associate with Highly Transcribed Gene Clusters

(A) Gene tracks represent, from the top, ERFS and ERFS hot spot demarcations; bindings of RPA, BRCA1, SMC5, γH2AX occupancy; and BrdU incorporation near the *IKZF1* locus. The y axis represents the total number of mapped reads per million of mapped reads (RPM) in 200 nucleotide windows (sliding-window smoothed).

(B) Genome-wide map of 619 ERFS hot spots. Each hot spot is represented by a green dot on the ideograms. The top fifteen hot spots are color-coded in red.

(C) Table of the top 15 ERFS hot spots. ERFS hot spots are ordered based on a ranked statistics of RPA/SMC5/BRCA1-binding strength (see Experimental Procedures). The first column depicts a representative gene within the hot spot. A hot spot containing at least three genes is designated as a "gene-cluster." A hot spot with a gene transcript value greater than 1 RPKM (reads per kilobase exon model per million mapped reads) is designated as transcribed. ERFS rearrangements in B cell cancers are listed in Table S2. ERFS is designated as "AID-target" according to (Chiarle et al., 2011; Klein et al., 2011). For complete definition of columns see Experimental Procedures. See also Tables S1 and S2.

Replication is organized into discrete zones (30–450 kbp in size) containing multiple replication origins that exhibit similar replication timing (Costa and Blow, 2007). Similarly, approximately 80% of the ERFSs are within 300 kbp of one another (Figures 2A and S4G). We therefore integrated these neighboring clustered ERFSs and removed those with footprints less than 10 kbp to define 619 triple colocalized hot spot regions (Figure 2B; Table S1). Interestingly, whereas these hot spots were distributed throughout the genome, the density of hot spots on autosomes was higher than on the sex chromosomes, which have a lower gene density (Figure 2B). An examination of the top 15 hot spots based on a ranked statistics of RPA/BRCA1/SMC5-binding strength showed that 9 out of the 15 regions contained gene clusters with at least three genes, and 12 out of 15 exhibited divergent/convergent gene pairs (Figures 2A and 2C; Table S1). Of note, 8 out of 15 hot spots are also rearranged in B cell lymphomas (Figure 2C; Table S2), suggesting a possible link among ERFSs, genome rearrangements, and cancer (see below).

Early S Phase Arrest by HU Induces DNA Damage at ERFSs, but Not at CFSs

DNA damage at CFSs is visualized by conventional cytogenetic analysis of metaphase chromosomes (Durkin and Glover, 2007). To investigate whether the ERFSs defined by RPA/ BRCA1/SMC5 binding are prone to actual breakage, we again treated cells with 10 mM HU, released them into fresh medium overnight, and examined metaphase spreads. Chromatid breaks, chromosome breaks, and rearrangements could be discerned in 20%–60% of WT cells after HU treatment (Figure S2D). To determine whether ERFSs are more sensitive to breakage under replication stress than regions lacking RPA/BRCA1/SMC5 binding (i.e., cold spots), we hybridized metaphases with bacterial artificial

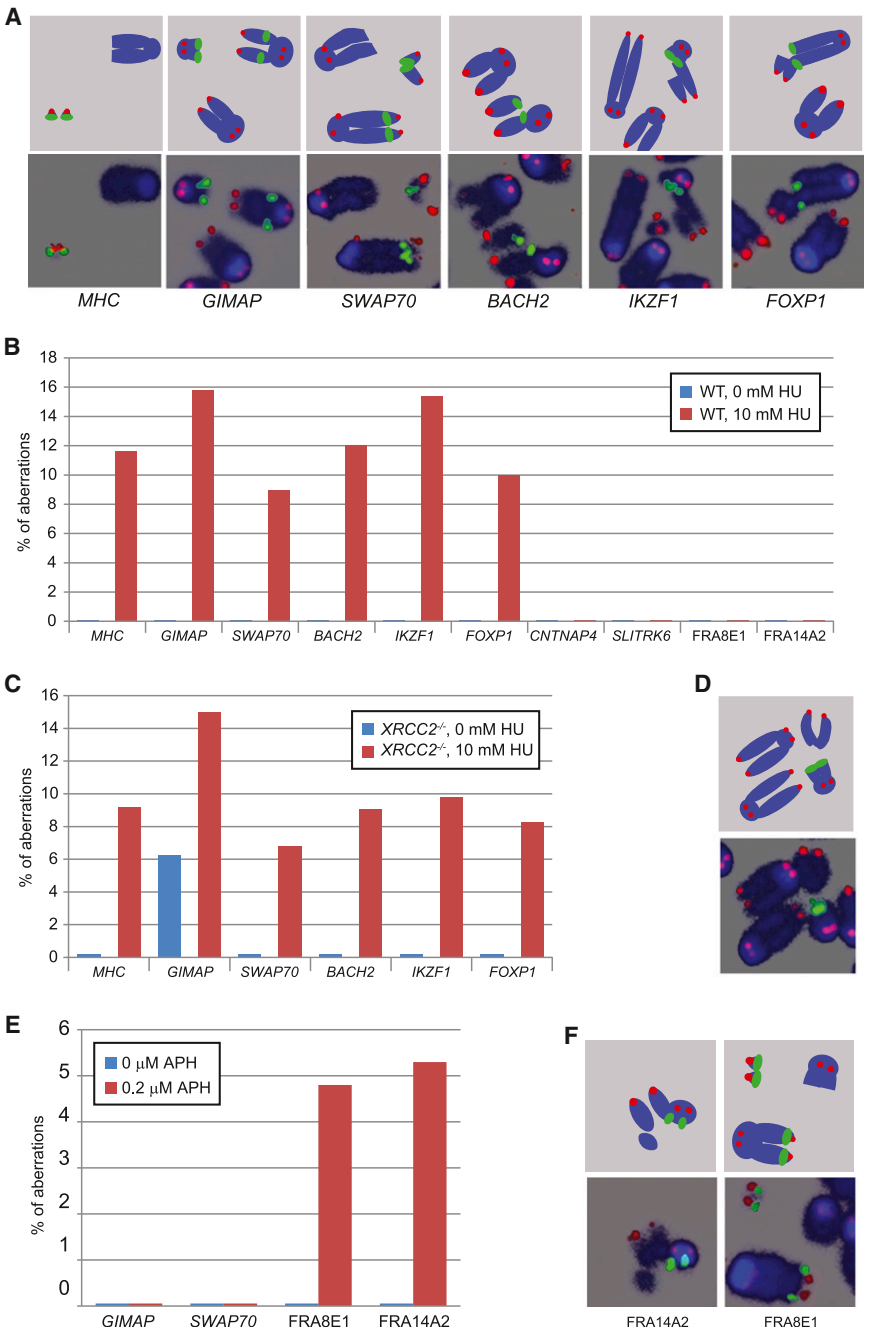


Figure 3. ERFS Break in Response to HU

(A) Upper: diagram of FISH probes. Lower: representative DNA aberrations identified by FISH. Blue is DAPI-stained DNA, green represents the BAC probe (*MHCII*, *GIMAP*, *SWAP70*, *BACH2*, *IKZF1*, or *FOXP1*) and red marks telomeric DNA.

(B) HU-induced aberrations were found at ERFSs but not at "cold sites" (*CNTNAP4*, *SLITRK6*) or CFSs (FRA8E1, FRA14A2). Quantitation of abnormalities from FISH analysis of untreated cells (blue bars) or cells treated with 10 mM HU (red bars). The percent aberrations specifically at the BAC probes relative to the total damage is plotted.

(C) Abnormalities detected by FISH in untreated (blue bars) and 10 mM HU-treated (red bars) *XRCC2*$^{-/-}$ cells.

(D) Upper: diagram of FISH probes. Lower: representative metaphase showing a spontaneous break at the *GIMAP* locus in an *XRCC2*$^{-/-}$ cell.

(E) Quantitation of abnormalities detected by FISH in untreated (blue bars) and 0.2 μM aphidicolin-treated (red bars) WT cells.

(F) Upper: diagram of FISH probes. Lower: representative metaphases showing aphidicolin-induced breaks at the FRA14A2 and FRA8E1 loci in WT cells. See also Figure S2 and Table S3.

chromosome (BAC) probes corresponding to six ERFS hot spots (*MHCII*, *GIMAP*, *SWAP70*, *BACH2*, *IKZF1*, and *FOXP1*) (Figures 3A and 3B), two cold spots (*CNTNAP4* and *SLITRK6*) and two CFSs (FRA8E1 and FRA14A2). For each of the six ERFS hot spots, a total of at least 40 chromosome aberrations were counted (Table S3). Notably, all six ERFS hot spots displayed chromosome aberrations in metaphases from HU-treated samples (Figure 3B). In contrast, neither of the cold regions or CFSs was broken under the same conditions (Figure 3B). Overall, 8%–15% of the total damage localized to individual ERFS hot spots, representing a significant fraction of the total damage (Figure 3B). DNA lesions were observed on either the centromeric or telomeric sides of ERFS-specific hybridized BAC (Figure S2E), suggesting that an ERFS represents a large fragile genomic region.

Aberrations at ERFS hot spots were also detected in *XRCC2*$^{-/-}$ cells treated with HU (Figure 3C). *XRCC2*$^{-/-}$ cells are more sensitive to HU than WT cells are, as evidenced by the higher level of total damage in these cells (Figure S2D). Breaks at *MHCII*, *GIMAP*, *SWAP70*, *BACH2*, *IKZF1*, and *FOXP1* were found in 5%–10% of HU-treated *XRCC2*$^{-/-}$ cells compared with 1%–6% of WT cells damaged in these regions (Table S3). Nevertheless, the frequency of ERFS-specific instability relative to the total damage was similar in *XRCC2*$^{-/-}$ and WT cells (Figures 3B and 3C). Interestingly, breaks in the vicinity of the *GIMAP* hot spot were detectable spontaneously in *XRCC2*$^{-/-}$ cells (Figures 3C and 3D; Table S3), which is consistent with increased γ-H2AX observed in unchallenged *XRCC2* mutant cells (Figure S2A).

None of the eight CFSs defined in mouse (Helmrich et al., 2006) were among our 619 ERFS hot spots (Table S1). Consistent with this, DNA aberrations at two of the most expressed CFSs in mouse lymphocytes, FRA14A2 and FRA8E1 (Helmrich et al., 2006) were undetectable in HU-treated WT samples

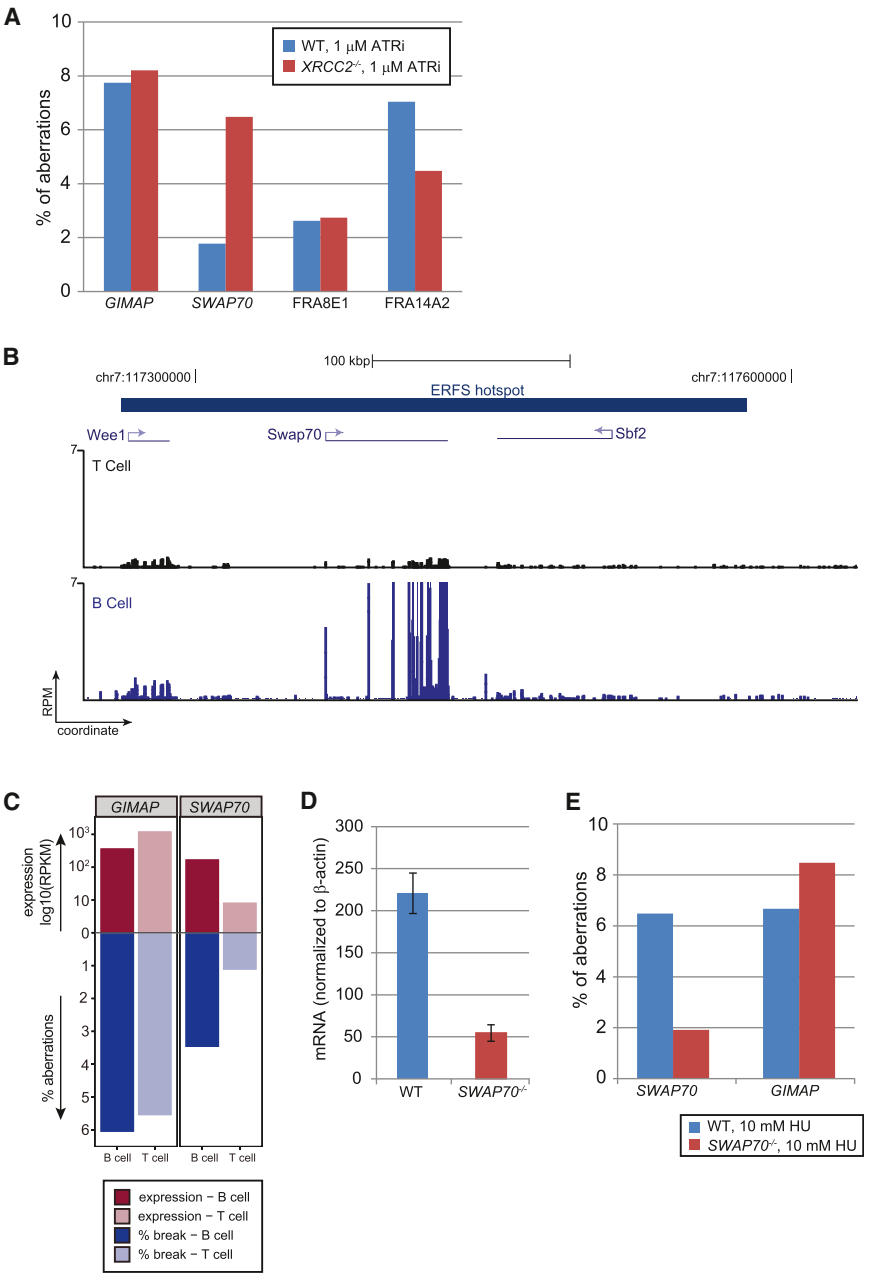


Figure 4. ERFS Break in Response to ATR Inhibition and High Transcription

(A) Quantitation of aberrations observed by FISH in response to overnight exposure to 1 μM ATRi in WT (blue bars) and *XRCC2*$^{-/-}$ cells (red bars).

(B) Gene tracks represent, from the top, ERFS demarcation and transcription measured by RNA-Seq in T and B cells at the region flanking *SWAP70* locus.

(C) Relative transcriptional activities of *GIMAP* and *SWAP70* loci in B and T cells and their relation to the ERFS fragility. *GIMAP* and SWAP70 hot spots are shown in separate facets. The x axis shows the cell lineage. The y axis upward depicts the log10(RPKM) in B and T cells by dark and light reds, respectively; the y axis downward depicts the quantitation of aberrations observed by FISH in response to overnight exposure to 1 μM ATRi in B and T cells in dark and light blue, respectively.

(D) Relative SWAP70 mRNA abundance (measured across exon 4) normalized to β-actin in WT and *SWAP70*$^{-/-}$ B cells (mean ± SD).

(E) Quantitation of aberrations in WT and *SWAP70*$^{-/-}$ cells at the *GIMAP* and *SWAP70* regions in response to 10 mM HU. See also Figure S5 and Table S3.

ATR Inhibition Promotes ERFS and CFS Expression

The ATR kinase protects the genome from chromosomal aberrations at late replicating CFSs, (Durkin and Glover, 2007) and is essential for stabilizing stalled forks and facilitates fork restart in early S phase (Cimprich and Cortez, 2008). To confirm that ATR inactivation induces CFSs and determine whether it similarly leads to damage at ERFSs, we treated asynchronous B cells on day 2 with 1 μM of a recently described ATR inhibitor (ATRi) (Toledo et al., 2011). We found that approximately 2.5% and 7.0% of the total chromosomal aberrations localized to the two CFSs, FRA8E1 and FRA14A2, respectively (Figure 4A). ATR deficiency also led to chromosomal aberrations at ERFSs at a similar frequency (Figure 4A; Table S3). Moreover, ERFSs and CFSs were both damaged in *XRCC2*$^{-/-}$ cells treated with ATRi (Figure 4A). Thus, the rupture of unreplicated regions at CFSs and fork collapse at ERFSs are similarly sensitive to ATR inhibition.

(Figure 3B). Absence of CFS expression could be explained by the fact that high concentrations of HU stall replication forks in early S phase (Figure 1A), whereas CFSs replicate late (Durkin and Glover, 2007). Conversely, we found that overnight treatment with low doses of aphidicolin (0.2 μM for 20 hr) induced damage at the CFSs FRA14A2 and FRA8E1, whereas the ERFSs *GIMAP* and *SWAP70* were largely insensitive (Figures 3E and 3F). These data are consistent with the idea that ERFS arise from fork collapse during early replication, whereas breakage at CFSs arises from a failure to replicate (Debatisse et al., 2012), and the two forms of replication stress induce distinct types of recurrent DNA lesions.

Transcriptional Activity Can Increase ERFS Fragility

As described above, ERFSs are enriched in regions with high transcriptional activity (Figures 1H, 2C, and S4F; Table S1). To determine the contribution of transcriptional activity to individual ERFSs, we focused on loci with tissue-specific transcription patterns. SWAP70 is a B-cell-specific developmental regulator, whereas genes within the *GIMAP* cluster are expressed both in

B and in T cells (Figures 4B and S5A). Treatment with ATRi led to a similar frequency of damage at *GIMAP* in B and T cells, consistent with insignificant changes in gene expression between the two cell types (Figure 4C). In contrast, damage near *SWAP70* was 3-fold lower in T than in B cells (Figure 4C; Table S3), which correlated with the decreased transcription of *SWAP70* in T cells (Figure 4B). Nevertheless, the replication timing near *SWAP70* was similar in both cell types (Figure S5B). To further delineate the role of transcription on ERFS breakage, we used *SWAP70*$^{-/-}$ mice in which 2.7 kbp, including the first exon and part of the 5′ untranslated region, is removed (Borggrefe et al., 2001), allowing us to compare the fragility of ERFSs in the same genomic region in knockout B cells. We determined that *SWAP70* mRNA in *SWAP70*$^{-/-}$ B cells was reduced by approximately 4-fold relative to levels in WT (Figure 4D). Moreover, DNA damage near *SWAP70* was approximately 2.5-fold lower in *SWAP70*$^{-/-}$ relative to levels in WT B cells (Figure 4E). In contrast, DNA damage near *GIMAP* remained at a similar level both in WT and *SWAP70*$^{-/-}$ cells (Figure 4E). Although our data indicate that high level of transcription contributes to the breakage of some ERFSs, other molecular features, including repetitive elements (Figure 1F), covalently bound protein complexes, and RNA:DNA hybrids, might also be sources of ERFS fragility.

Oncogenic Stress Can Trigger ERFS and CFS Fragility

Oncogene deregulation is thought to compromise genome integrity preferentially at CFSs (Bartek et al., 2007; Halazonetis et al., 2008), and CFS deletion has been associated with various cancers (Bignell et al., 2010). To determine whether oncogenic stress similarly induces DNA damage at ERFSs, we overexpressed *c-myc* in B cells because it has been implicated in regulating replication initiation and origin firing (Dominguez-Sola et al., 2007). *XRCC2*$^{-/-}$ cells were utilized to increase the amount of replicative stress and DNA damage as a result of decreased HR (Figure S2D). *c-myc* overexpression led to induction of p53 (Figure 5A), which correlated with an approximately 1.6-fold increase in overall DNA damage in *XRCC2*$^{-/-}$ cells overexpressing *c-myc* compared to empty vector (EV)-infected cells (Table S3). Moreover, 7.3% of the total breaks generated in *c-myc* overexpressing cells were found near *SWAP70*, compared to 2.4% of total breaks at this ERFS in EV-infected B cells (Figure 5B). Similarly, out of 43 breaks observed in *c-myc*-infected cells, 3 (7%) were found at the *GIMAP* cluster, and 3 (6.7%) were found near BACH2. *c-myc* overexpression also induced breaks at FRA8E1, showing a 2-fold relative increase in breaks relative to EV-infected cells (Figure 5B). Thus, DNA damage induced by *c-myc* overexpression can occur at ERFSs and CFSs.

ERFS Fragility Is AID Independent

Mutations and DSBs at various oncogenes, including *c-myc*, are due to AID off-target activity (Robbiani et al., 2008). Recently, a number of genome-wide studies in primary B cells mapped AID-induced DNA translocation events, and identified several novel hot spots for AID-dependent translocations at non-Ig genes (Chiarle et al., 2011; Kato et al., 2012; Klein et al., 2011). Among these translocation hot spots, *MHCII*, *GIMAP*, *IKZF1*, *PVT1*, *ETS1*, *IRF4*, and *NFkB1* were located within the top 15 ERFS hot spots in this study, whereas the *IgH* locus (the physiologic target of AID) was not ranked high on the list (Figure 2C; Table S1). To determine whether AID contributes to ERFS fragility, we stimulated WT and AID knockout B cells with LPS/IL4 for 2 days, and then treated them with ATRi overnight. These conditions induce robust AID-dependent DNA damage simultaneously with replication stress. We probed metaphases with BACs spanning the *IgH* locus, the *GIMAP* cluster, and *IKZF1*-all AID translocation hot spots—as well as *BACH2*, *SWAP70*, *FOXP1*, and *BCL2* (Figure S2E)—ERFSs that are frequently rearranged in B cell lymphoma (Figure 2C; Tables S1 and S2). In WT, the IgH locus was damaged in 3.8% of cells, but the frequency of *IgH*-specific instability did not increase with ATRi (Figure S2F), despite the fact that ATRi greatly increased overall damage (Table S3). Upon ATRi treatment, the frequency of breaks at the ERFSs *GIMAP*, *IKZF1*, *BACH2*, *SWAP70*, and *FOXP1*, and *BCL2* were elevated to the levels similar to those observed at the *IgH* in activated B cells (Figure S2F). Breaks at some ERFSs were even spontaneously detected (*FOXP1* and *GIMAP*, Figure S2F).

To determine whether AID expression contributes to aberrations observed at ERFSs, we next analyzed their breakage frequency in *AID*$^{-/-}$ cells. Unlike WT cells, *IgH* breaks were absent in *AID*$^{-/-}$ cells. In contrast, all ERFSs exhibited similar levels of breakage both in WT and *AID*$^{-/-}$ cells (Figure 5C; Table S3). Therefore, whereas *IgH* breaks in B cells are entirely AID dependent, the breakage of ERFSs is AID independent. Altogether, these data suggest that some recurrent rearrangements in B cell lymphoma are due to AID-independent replicative stress at ERFSs.

Genome Instability at ERFSs Is Observed in Mouse Models and Human Cancer

Among the top 15 ERFS hot spot that break in response to AID-independent replication stress, we have identified three partners that recurrently translocate to *IgH* in lymphomas: *BACH2*, *FOXP*1, and *BCL2* (Table S2). We hypothesized that if AID-dependent DSBs in G1 persisted into early S phase, translocations between AID-dependent breaks and ERFS might be detectable. To test this, we examined cells transgenically overexpressing AID and simultaneously deficient for 53BP1 (IgκAID/*53BP1*$^{-/-}$), thus allowing the persistence of G1 *IgH* breaks into S phase where they could be joined to ERFSs. Indeed, 26% and 7% of IgκAID/*53BP1*$^{-/-}$ B cells carried *IgH* locus and *BACH2* breaks, respectively (Figure 5D). These breaks are fusogenic because *IgH*- and *BACH2*-associated translocations to unidentified partner chromosomes were found in 7.3% and 1.2% of the metaphases, respectively (Figure 5D). Importantly, we also detected one *IgH*/*BACH2* translocation among 750 cells (Figure 5E), reminiscent of the IgH/BACH2 translocations observed in human B cell lymphoma (Kobayashi et al., 2011). Thus, AID-dependent breaks generated in G1 (Petersen et al., 2001) can join to ERFS breaks triggered in early S phase.

A hallmark of cancer genomes is widespread copy-number changes, insertions, and deletions. To determine whether deletions and/or amplifications at ERFSs are a general feature of the B cell lymphoma genome, we compared our ERFSs with

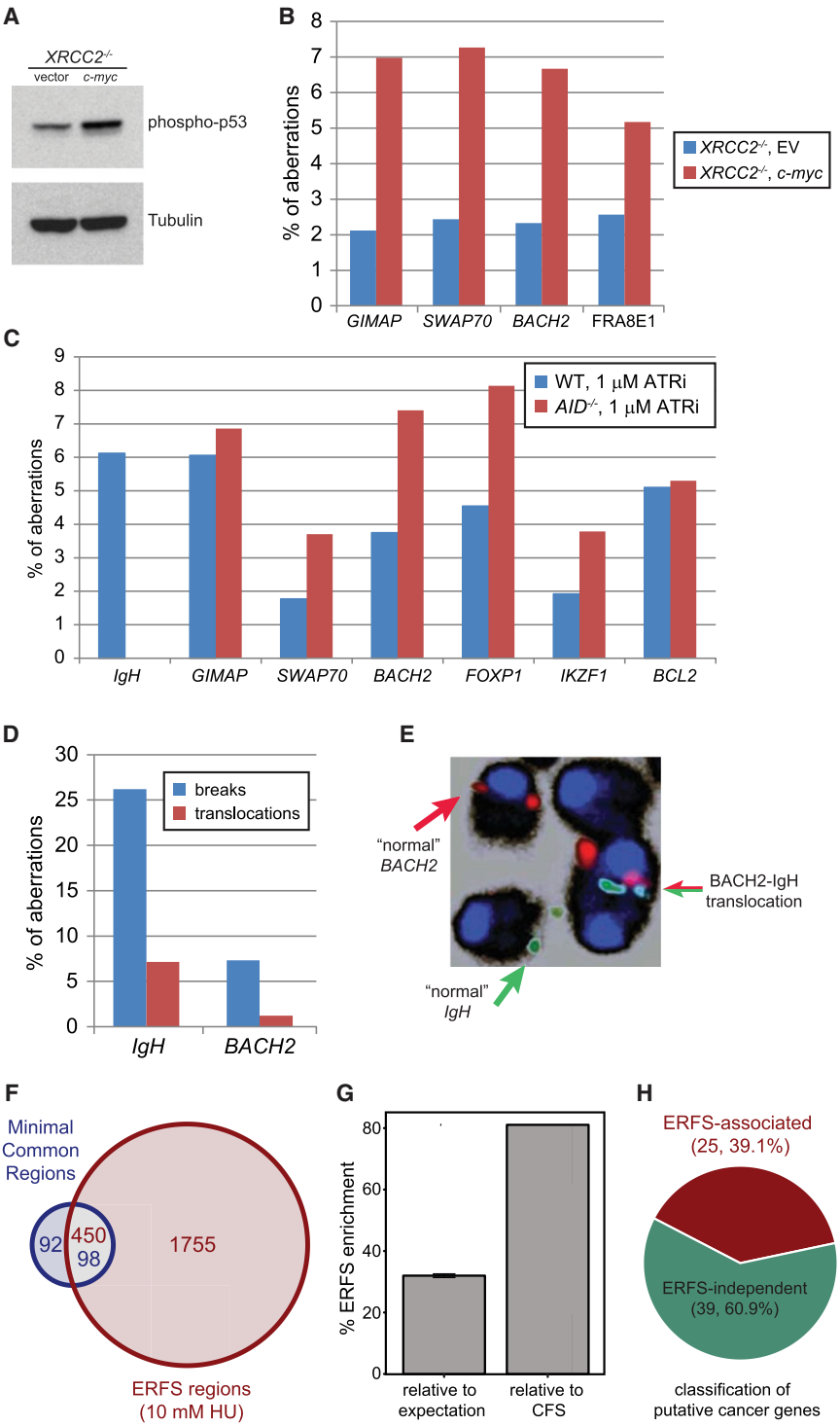


Figure 5. ERFS Fragility Is Observed in Response to Oncogenic Stress and in Human Cancer

(A) Western blot for phosphorylated p53 in *c-myc* and EV-infected *XRCC2*$^{-/-}$ B cells.

(B) Aberrations in *c-myc*-infected and EV-infected *XRCC2*$^{-/-}$ B cells.

(C) Aberrations in WT (blue bars) and *AID*$^{-/-}$ B cells (red bars) treated with 1 μM ATRi.

(D) Spontaneous chromosome breaks (blue bars) and translocations (red bars) at the *IgH* and *BACH2* locus in IgκAID/*53BP1*$^{-/-}$ B cells.

(E) Normal chromosomes and a translocation of *BACH2* ERFS (red) to the *IgH* locus (green) is shown.

(F and G) ERFSs significantly overlap with MCRs detected in DLBCL. The Venn diagram shows the overlap of ERFSs with MCR found in DLBCL. The total number of regions is indicated for each shared and unique area and color-coded based on the region's title.

(G) Significance of correlation between the ERFSs and MCRs is evaluated relative to the permutation model and CFSs. The percent increase in the overlap between the ERFSs and MCRs relative to the permutation model's expectation (mean ± SEM, $p < 1 \times 10^{-4}$) and CFSs are shown in the left and right bar graphs, respectively.

(H) ERFSs are enriched for known cancer genes. The pie chart shows the fraction of putative cancer genes (Bignell et al., 2010) associated with ERFSs ($p < 6 \times 10^{-20}$). See also Figure S6 and Tables S3 and S4.

high resolution copy-number changes detected in biopsies of patients with diffuse large B cell lymphoma (DLBCL), the most common type of non-Hodgkins lymphoma (Lenz et al., 2008). A total of 190 "minimal common regions" (MCRs) were found among 203 biopsies, carrying a gain or a loss of a chromosomal region ranging in size from 5 kbp to 21Mbp (Lenz et al., 2008). Mouse ERFS coordinates were overlaid onto the human genome using two methods, yielding 2,205 syntenic regions (Figures S6B–S6D). Notably, 51.6% of the MCRs observed in primary DLBCL overlapped with syntenic ERFS regions (p(permutation) $< 1 \times 10^{-4}$, Figure 5F). Moreover, 20.4% of ERFSs overlapped with MCRs, 32% higher than expectation (p(permutation) $< 1 \times 10^{-6}$, Figure 5G). Surprisingly, ERFS were deleted or amplified in DLBCL at least 81% more frequently as compared to CFSs, despite their cancer-specific propensity for breakage (Figure 5G). Moreover, our analysis indicated that the DLBCL copy-number alterations exhibited 2-fold higher correlation with B cell ERFSs compared to deletions and/or amplifications in T lineage acute lymphoblastic leukemia (Figure S6A) (Zhang et al., 2012a). Finally, by examining homozygous deletions in cancer genomes (Bignell et al., 2010), we found that 25 out of 64 genes known to contribute to oncogenesis coincide with ERFSs (p(hypergeometric) $< 6 \times 10^{-20}$, Figure 5H; Table S4). Based on these findings,

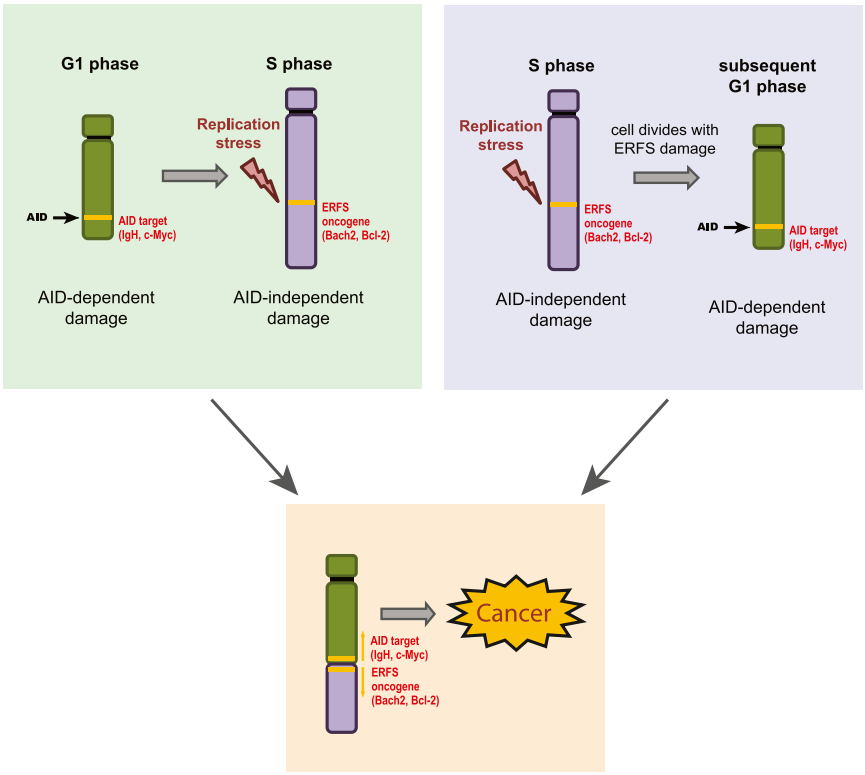


Figure 6. Model for Recurrent Rearrangements in B Cell Lymphomas

AID is active in G1 (Petersen et al., 2001) and targets IgH and various oncogenes (e.g., *c-myc*). Replication fork collapse at ERFSs in S phase occurs at preferential sites including various cancer-associated genes (e.g., *BCL2*, *BACH2*). An AID-generated break might be passed from G1 to early S, where it meets an ERFS, which may eventually result in a translocation (left). Alternatively, an ERFS (bearing unresolved a replication intermediate of under-replicated DNA) might break in mitosis and then become permissive to translocate to an AID-induced DSB in the next G1 phase of the cell cycle (right).

we conclude that ERFSs are a significant feature of the mutational landscape of diffuse large B cell lymphomas and potentially other cancers.

DISCUSSION

Replicative Stress at ERFSs Contributes to Genome Instability in B Cells

Although AID has been implicated in B cell translocations (Gostissa et al., 2011), very little is known about the mechanisms of chromosomal breakage at several *IgH*-partner loci, including *BCL2*, *BACH2*, and *FOXP1*. Besides programmed DNA damage, replication-based mechanisms are a major contributor to chromosomal instability in cancer (Liu et al., 2012). Activated B cells are among the most rapidly dividing mammalian cells (Zhang et al., 1988), which potentially exposes them to high endogenous levels of replicative stress. Here, we have used a genome-wide approach to identify a subset of early replicating regions in the B cell genome that are particularly vulnerable to fork collapse and contribute to rearrangements in B cell malignancies. In our model, ERFS breaks can occur after the generation of unrepaired AID-induced breaks in G1, and the two breaks could recombine during S or G2. Alternatively, ERFS damage might persist through mitosis resulting in DNA breaks in the subsequent G1 phase when AID is predominantly active. In either case, we suggest that AID-mediated DSBs in G1 (Petersen et al., 2001), together with replication-stress-induced damage at recurrent loci, can coordinately drive B cell lymphoma initiation and progression (Figure 6).

ERFS versus CFS

CFSs are considered to be the most replication-stress-sensitive sites in the genome (Durkin and Glover, 2007). Although no single mechanism accounts for CFS instability, it is hypothesized that a number of different characteristics may contribute to their fragility including co-occurrence with very large genes, late replication, low density of replication origins, high A-T content, and sequences prone to form secondary structures, histone hypoacetylation, and a condensed chromatin structure (Helmrich et al., 2011; Jiang et al., 2009; Letessier et al., 2011; Ozeri-Galai et al., 2011). In stark contrast to CFSs, our identified ERFSs replicate early; have an open chromatin configuration; and are origin-, gene-, and G-C-rich.

Despite these diametrically opposite properties, both CFS and ERFS fragility are increased by ATR inhibition (Figure 4A), oncogenic stress (Figure 5B), and deficiencies in HR (Figure 3C) (Bartek et al., 2007; Durkin and Glover, 2007; Halazonetis et al., 2008). These conditions decrease the rate of fork progression but concomitantly increase the density of replication initiating events (Bester et al., 2011; Daboussi et al., 2008; Dominguez-Sola et al., 2007; Shechter et al., 2004), which might contribute to the damage at both CFSs and ERFSs, respectively. The decrease in fork speed hinders the completion of replication at CFSs, either because of the scarcity of origins near CFSs (Letessier et al., 2011), the heterochromatic nature of the regions that would limit accessibility of DNA replication and/or DSB repair machineries (Jiang et al., 2009), or because of the interference between transcription and replication at very large genes (Helmrich et al., 2011). Although additional origins are not activated near CFSs upon replication stress (Letessier et al., 2011), an increase in origin activity at early replicons might paradoxically contribute to genome instability at ERFSs. For example, increasing the replication initiation events near highly transcribed gene clusters with divergent and/or convergent gene pairs could increase conflicts between DNA replication and transcription machineries. The higher density of activated origins at ERFSs would also be expected to prematurely deplete

nucleotide pools (Bester et al., 2011), thereby increasing the probability of subsequent fork stalling and collapse. These two outcomes of replication stress are likely to be linked because increased replication initiation and depletion of nucleotide supplies slows replication (Bester et al., 2011; Jones et al., 2012), whereas slow fork progression causes activation of dormant origins (Ge et al., 2007), and both incomplete replication and increased origin firing are monitored by ATR activity (Shechter et al., 2004). In conclusion, increased initiating events at ERFSs and a paucity of replication initiation at CFSs could both challenge replication fidelity.

ERFSs and Cancer

Oncogenic stress is a major driving force in the early stages of cancer development (Halazonetis et al., 2008); nevertheless, the factors that trigger replicative stress in vivo remain unclear. In the case of B cell lymphomas, oncogenic stress can be initiated by the activity of AID, which by targeting non-*Ig* genes such as *c-myc* (Robbiani et al., 2008), leads to *c-myc*/*IgH* translocations and consequent aberrant *c-myc* expression. This form of AID-induced oncogenic stress or high levels of proliferative activity in activated B cells could generate DNA damage at ERFSs (Figure 6).

Altogether, 103 AID hot spots (Chiarle et al., 2011; Klein et al., 2011)—including the *GIMAP* cluster, *MHCII* locus, and *IKZF1*—were also identified as ERFS hot spots in this study (Table S1). It is possible that the overlap observed between a subset of off-target AID sites and ERFSs is due to common underlying features of these loci. For example, AID is recruited to ssDNA regions (Chaudhuri and Alt, 2004), which are also generated during replicative stress; AID-dependent DSBs and ERFSs are also both enriched in repeat elements (Staszewski et al., 2011). In addition, chromosomal regions with the highest transcriptional activity have the highest AID-dependent translocation density (Chiarle et al., 2011; Klein et al., 2011), and early origins and translocations frequently reside near transcription start sites and RNA polymerase-II-binding sites (Chiarle et al., 2011; Klein et al., 2011). Thus, these euchromatic regions could serve both as AID targets in G1 and also be susceptible to fork collapse during early S phase.

A number of different hypotheses have been put forward about the mechanisms that promote recurrent translocations in mature B cell lymphomas. These include recurrent genomic damage by AID, random DNA damage followed by selection, and a nonrandom 3D organization of the genome (Chiarle et al., 2011; Hakim et al., 2012; Klein et al., 2011; Zhang et al., 2012b). To date, replication-stress-induced DNA damage has been associated with late-replicating CFS. By using an alternative experimental approach for the discovery of fragile site expression during early replication, we have identified a novel source of recurrent AID-independent DNA breaks that may play a mechanistic role in some of the most common genome rearrangements during B cell lymphomagenesis. Because transcriptional activity and replication timing of a genomic region vary among different cell lineages (Hansen et al., 2010), different sets of ERFSs might also account for recurrent chromosomal rearrangements in cancers of distinct cellular origins.

EXPERIMENTAL PROCEDURES

Mice

XRCC2$^{-/-}$(Frappart et al., 2009), 53BP1$^{-/-}$ (Ward et al., 2004), IgκAID (Robbiani et al., 2009), *AID*$^{-/-}$ (Muramatsu et al., 2000), and *SWAP70*$^{-/-}$ (Borggrefe et al., 2001) mice have been described. *SWAP70*$^{-/-}$ and WT control mice used in Figure 4D are C57BL/6 background; all other mice are 129/Sv x C57BL/6 background.

ChIP-Seq, Repli-Seq, RNA-Seq, DHS I Mapping, and FISH Analysis

ChIP-seq and RNA-Seq procedures were performed as in Yamane et al. (2011), Repli-Seq was performed as described in Hansen et al. (2010). DHSI mapping was performed as described (Sekimata et al., 2009), and fluorescence in situ hybridization (FISH) analysis is described in Callén et al. (2007). For detailed methods, see Extended Experimental Procedures.

BACs

Individual BACs to ERFSs were identified using NCBI clone finder and purchased from BACPAC. For complete list of BAC probes used in FISH experiments see Extended Experimental Procedures.

Retroviral Infection

Cells were infected with pMX-c-Myc-IRES-GFP or empty vector and GFP-positive cells were sorted as described (Robbiani et al., 2008).

Statistical and Computational Analyses

Detailed description is available in Extended Experimental Procedures.

ACCESSION NUMBERS

The ChIP-seq and RNA-seq data are deposited in GEO under accession number GSE43504.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, six figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2013.01.006.

ACKNOWLEDGMENTS

We thank Rolf Jessberger and Alessandra Pernis for *SWAP70*$^{-/-}$ mice, Mila Jankovic for pMX-c-Myc-IRES-GFP, Jean Gautier, Mirit Aladjem, and Toren Finkel for discussions and Lars Grontved for help with DHS mapping. This work was supported by the Intramural Research Program of the NIH, the National Cancer Institute, and the Center for Cancer Research, and by a Department of Defense grant to A.N. (BC102335). This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the NIH.

Received: November 9, 2012
Revised: December 10, 2012
Accepted: January 2, 2013
Published: January 24, 2013

REFERENCES

Bartek, J., Bartkova, J., and Lukas, J. (2007). DNA damage signalling guards against activated oncogenes and tumour progression. Oncogene *26*, 7773–7779.

Bester, A.C., Roniger, M., Oren, Y.S., Im, M.M., Sarni, D., Chaoat, M., Bensimon, A., Zamir, G., Shewach, D.S., and Kerem, B. (2011). Nucleotide deficiency promotes genomic instability in early stages of cancer development. Cell *145*, 435–446.

Bignell, G.R., Greenman, C.D., Davies, H., Butler, A.P., Edkins, S., Andrews, J.M., Buck, G., Chen, L., Beare, D., Latimer, C., et al. (2010). Signatures of mutation and selection in the cancer genome. Nature *463*, 893–898.

Borggrefe, T., Keshavarzi, S., Gross, B., Wabl, M., and Jessberger, R. (2001). Impaired IgE response in SWAP-70-deficient mice. Eur. J. Immunol. *31*, 2467–2475.

Bothmer, A., Robbiani, D.F., Feldhahn, N., Gazumyan, A., Nussenzweig, A., and Nussenzweig, M.C. (2010). 53BP1 regulates DNA resection and the choice between classical and alternative end joining during class switch recombination. J. Exp. Med. *207*, 855–865.

Bunting, S.F., Callén, E., Kozak, M.L., Kim, J.M., Wong, N., López-Contreras, A.J., Ludwig, T., Baer, R., Faryabi, R.B., Malhowski, A., et al. (2012). Mol. Cell *46*, 125–135.

Bunting, S.F., Callén, E., Wong, N., Chen, H.T., Polato, F., Gunn, A., Bothmer, A., Feldhahn, N., Fernandez-Capetillo, O., Cao, L., et al. (2010). 53BP1 inhibits homologous recombination in Brca1-deficient cells by blocking resection of DNA breaks. Cell *141*, 243–254.

Callén, E., Jankovic, M., Difilippantonio, S., Daniel, J.A., Chen, H.T., Celeste, A., Pellegrini, M., McBride, K., Wangsa, D., Bredemeyer, A.L., et al. (2007). ATM prevents the persistence and propagation of chromosome breaks in lymphocytes. Cell *130*, 63–75.

Cha, R.S., and Kleckner, N. (2002). ATR homolog Mec1 promotes fork progression, thus averting breaks in replication slow zones. Science *297*, 602–606.

Chaudhuri, J., and Alt, F.W. (2004). Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. Nat. Rev. Immunol. *4*, 541–552.

Chen, H.T., Bhandoola, A., Difilippantonio, M.J., Zhu, J., Brown, M.J., Tai, X., Rogakou, E.P., Brotz, T.M., Bonner, W.M., Ried, T., and Nussenzweig, A. (2000). Response to RAG-mediated VDJ cleavage by NBS1 and gamma-H2AX. Science *290*, 1962–1965.

Chiarle, R., Zhang, Y., Frock, R.L., Lewis, S.M., Molinie, B., Ho, Y.J., Myers, D.R., Choi, V.W., Compagno, M., Malkin, D.J., et al. (2011). Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell *147*, 107–119.

Cimprich, K.A., and Cortez, D. (2008). ATR: an essential regulator of genome integrity. Nat. Rev. Mol. Cell Biol. *9*, 616–627.

Costa, S., and Blow, J.J. (2007). The elusive determinants of replication origins. EMBO Rep. *8*, 332–334.

Daboussi, F., Courbet, S., Benhamou, S., Kannouche, P., Zdzienicka, M.Z., Debatisse, M., and Lopez, B.S. (2008). A homologous recombination defect affects replication-fork progression in mammalian cells. J. Cell Sci. *121*, 162–166.

Debatisse, M., Le Tallec, B., Letessier, A., Dutrillaux, B., and Brison, O. (2012). Common fragile sites: mechanisms of instability revisited. Trends Genet. *28*, 22–32.

Dominguez-Sola, D., Ying, C.Y., Grandori, C., Ruggiero, L., Chen, B., Li, M., Galloway, D.A., Gu, W., Gautier, J., and Dalla-Favera, R. (2007). Non-transcriptional control of DNA replication by c-Myc. Nature *448*, 445–451.

Durkin, S.G., and Glover, T.W. (2007). Chromosome fragile sites. Annu. Rev. Genet. *41*, 169–192.

Feng, W., Collingwood, D., Boeck, M.E., Fox, L.A., Alvino, G.M., Fangman, W.L., Raghuraman, M.K., and Brewer, B.J. (2006). Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. Nat. Cell Biol. *8*, 148–155.

Frappart, P.O., Lee, Y., Russell, H.R., Chalhoub, N., Wang, Y.D., Orii, K.E., Zhao, J., Kondo, N., Baker, S.J., and McKinnon, P.J. (2009). Recurrent genomic alterations characterize medulloblastoma arising from DNA double-strand break repair deficiency. Proc. Natl. Acad. Sci. USA *106*, 1880–1885.

Ge, X.Q., Jackson, D.A., and Blow, J.J. (2007). Dormant origins licensed by excess Mcm2-7 are required for human cells to survive replicative stress. Genes Dev. *21*, 3331–3341.

Gostissa, M., Alt, F.W., and Chiarle, R. (2011). Mechanisms that promote and suppress chromosomal translocations in lymphocytes. Annu. Rev. Immunol. *29*, 319–350.

Hakim, O., Resch, W., Yamane, A., Klein, I., Kieffer-Kwon, K.R., Jankovic, M., Oliveira, T., Bothmer, A., Voss, T.C., Ansarah-Sobrinho, C., et al. (2012). DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. Nature *484*, 69–74.

Halazonetis, T.D., Gorgoulis, V.G., and Bartek, J. (2008). An oncogene-induced DNA damage model for cancer development. Science *319*, 1352–1355.

Hansen, R.S., Thomas, S., Sandstrom, R., Canfield, T.K., Thurman, R.E., Weaver, M., Dorschner, M.O., Gartler, S.M., and Stamatoyannopoulos, J.A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc. Natl. Acad. Sci. USA *107*, 139–144.

Hashash, N., Johnson, A.L., and Cha, R.S. (2011). Regulation of fragile sites expression in budding yeast by MEC1, RRM3 and hydroxyurea. J. Cell Sci. *124*, 181–185.

Helmrich, A., Stout-Weider, K., Hermann, K., Schrock, E., and Heiden, T. (2006). Common fragile sites are conserved features of human and mouse chromosomes and relate to large active genes. Genome Res. *16*, 1222–1230.

Helmrich, A., Ballarino, M., and Tora, L. (2011). Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. Mol. Cell *44*, 966–977.

Jiang, Y., Lucas, I., Young, D.J., Davis, E.M., Karrison, T., Rest, J.S., and Le Beau, M.M. (2009). Common fragile sites are characterized by histone hypoacetylation. Hum. Mol. Genet. *18*, 4501–4512.

Jones, R.M., Mortusewicz, O., Afzal, I., Lorvellec, M., García, P., Helleday, T., and Petermann, E. (2012). Increased replication initiation and conflicts with transcription underlie Cyclin E-induced replication stress. Oncogene. Published online September 3, 2012.

Kato, L., Begum, N.A., Burroughs, A.M., Doi, T., Kawai, J., Daub, C.O., Kawaguchi, T., Matsuda, F., Hayashizaki, Y., and Honjo, T. (2012). Nonimmunoglobulin target loci of activation-induced cytidine deaminase (AID) share unique features with immunoglobulin genes. Proc. Natl. Acad. Sci. USA *109*, 2479–2484.

Klein, I.A., Resch, W., Jankovic, M., Oliveira, T., Yamane, A., Nakahashi, H., Di Virgilio, M., Bothmer, A., Nussenzweig, A., Robbiani, D.F., et al. (2011). Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. Cell *147*, 95–106.

Kobayashi, S., Taki, T., Chinen, Y., Tsutsumi, Y., Ohshiro, M., Kobayashi, T., Matsumoto, Y., Kuroda, J., Horiike, S., Nishida, K., and Taniwaki, M. (2011). Identification of IGHCδ-BACH2 fusion transcripts resulting from cryptic chromosomal rearrangements of 14q32 with 6q15 in aggressive B-cell lymphoma/leukemia. Genes Chromosomes Cancer *50*, 207–216.

Lenz, G., Wright, G.W., Emre, N.C., Kohlhammer, H., Dave, S.S., Davis, R.E., Carty, S., Lam, L.T., Shaffer, A.L., Xiao, W., et al. (2008). Molecular subtypes of diffuse large B-cell lymphoma arise by distinct genetic pathways. Proc. Natl. Acad. Sci. USA *105*, 13520–13525.

Letessier, A., Millot, G.A., Koundrioukoff, S., Lachagès, A.M., Vogt, N., Hansen, R.S., Malfoy, B., Brison, O., and Debatisse, M. (2011). Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. Nature *470*, 120–123.

Liu, M., Duke, J.L., Richter, D.J., Vinuesa, C.G., Goodnow, C.C., Kleinstein, S.H., and Schatz, D.G. (2008). Two levels of protection for the B cell genome during somatic hypermutation. Nature *451*, 841–845.

Liu, P., Carvalho, C.M., Hastings, P.J., and Lupski, J.R. (2012). Mechanisms for recurrent and complex human genomic rearrangements. Curr. Opin. Genet. Dev. *22*, 211–220.

MacAlpine, D.M., Rodríguez, H.K., and Bell, S.P. (2004). Coordination of replication and transcription along a Drosophila chromosome. Genes Dev. *18*, 3094–3105.

Mirkin, E.V., and Mirkin, S.M. (2007). Replication fork stalling at natural impediments. Microbiol. Mol. Biol. Rev. *71*, 13–35.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. Cell *102*, 553–563.

Ozeri-Galai, E., Lebofsky, R., Rahat, A., Bester, A.C., Bensimon, A., and Kerem, B. (2011). Failure of origin activation in response to fork stalling leads to chromosomal instability at fragile sites. Mol. Cell *43*, 122–131.

Petermann, E., Orta, M.L., Issaeva, N., Schultz, N., and Helleday, T. (2010). Hydroxyurea-stalled replication forks become progressively inactivated and require two different RAD51-mediated pathways for restart and repair. Mol. Cell *37*, 492–502.

Petersen, S., Casellas, R., Reina-San-Martin, B., Chen, H.T., Difilippantonio, M.J., Wilson, P.C., Hanitsch, L., Celeste, A., Muramatsu, M., Pilch, D.R., et al. (2001). AID is required to initiate Nbs1/gamma-H2AX focus formation and mutations at sites of class switching. Nature *414*, 660–665.

Raveendranathan, M., Chattopadhyay, S., Bolon, Y.T., Haworth, J., Clarke, D.J., and Bielinsky, A.K. (2006). Genome-wide replication profiles of S-phase checkpoint mutants reveal fragile sites in yeast. EMBO J. *25*, 3627–3639.

Robbiani, D.F., Bothmer, A., Callen, E., Reina-San-Martin, B., Dorsett, Y., Difilippantonio, S., Bolland, D.J., Chen, H.T., Corcoran, A.E., Nussenzweig, A., and Nussenzweig, M.C. (2008). AID is required for the chromosomal breaks in c-myc that lead to c-myc/IgH translocations. Cell *135*, 1028–1038.

Robbiani, D.F., Bunting, S., Feldhahn, N., Bothmer, A., Camps, J., Deroubaix, S., McBride, K.M., Klein, I.A., Stone, G., Eisenreich, T.R., et al. (2009). AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. Mol. Cell *36*, 631–641.

Schlacher, K., Wu, H., and Jasin, M. (2012). A distinct replication fork protection pathway connects Fanconi anemia tumor suppressors to RAD51-BRCA1/2. Cancer Cell *22*, 106–116.

Sekimata, M., Pérez-Melgosa, M., Miller, S.A., Weinmann, A.S., Sabo, P.J., Sandstrom, R., Dorschner, M.O., Stamatoyannopoulos, J.A., and Wilson, C.B. (2009). CCCTC-binding factor and the transcription factor T-bet orchestrate T helper 1 cell-specific structure and function at the interferon-gamma locus. Immunity *31*, 551–564.

Shechter, D., Costanzo, V., and Gautier, J. (2004). ATR and ATM regulate the timing of DNA replication origin firing. Nat. Cell Biol. *6*, 648–655.

Sonoda, E., Sasaki, M.S., Buerstedde, J.M., Bezzubova, O., Shinohara, A., Ogawa, H., Takata, M., Yamaguchi-Iwai, Y., and Takeda, S. (1998). Rad51-deficient vertebrate cells accumulate chromosomal breaks prior to cell death. EMBO J. *17*, 598–608.

Stamatoyannopoulos, J.A., Snyder, M., Hardison, R., Ren, B., Gingeras, T., Gilbert, D.M., Groudine, M., Bender, M., Kaul, R., Canfield, T., et al.; Mouse ENCODE Consortium. (2012). An encyclopedia of mouse DNA elements (Mouse ENCODE). Genome Biol. *13*, 418.

Staszewski, O., Baker, R.E., Ucher, A.J., Martier, R., Stavnezer, J., and Guikema, J.E. (2011). Activation-induced cytidine deaminase induces reproducible DNA breaks at many non-Ig Loci in activated B cells. Mol. Cell *41*, 232–242.

Stephan, A.K., Kliszczak, M., and Morrison, C.G. (2011). The Nse2/Mms21 SUMO ligase of the Smc5/6 complex in the maintenance of genome stability. FEBS Lett. *585*, 2907–2913.

Tanaka, T., and Nasmyth, K. (1998). Association of RPA with chromosomal replication origins requires an Mcm protein, and is regulated by Rad53, and cyclin- and Dbf4-dependent kinases. EMBO J. *17*, 5182–5191.

Toledo, L.I., Murga, M., Zur, R., Soria, R., Rodriguez, A., Martinez, S., Oyarzabal, J., Pastor, J., Bischoff, J.R., and Fernandez-Capetillo, O. (2011). A cell-based screen identifies ATR inhibitors with synthetic lethal properties for cancer-associated mutations. Nat. Struct. Mol. Biol. *18*, 721–727.

Tsai, A.G., Lu, H., Raghavan, S.C., Muschen, M., Hsieh, C.L., and Lieber, M.R. (2008). Human chromosomal translocations at CpG sites and a theoretical basis for their lineage and stage specificity. Cell *135*, 1130–1142.

Ward, I.M., and Chen, J. (2001). Histone H2AX is phosphorylated in an ATR-dependent manner in response to replicational stress. J. Biol. Chem. *276*, 47759–47762.

Ward, I.M., Reina-San-Martin, B., Olaru, A., Minn, K., Tamada, K., Lau, J.S., Cascalho, M., Chen, L., Nussenzweig, A., Livak, F., et al. (2004). 53BP1 is required for class switch recombination. J. Cell Biol. *165*, 459–464.

Yamane, A., Resch, W., Kuo, N., Kuchen, S., Li, Z., Sun, H.W., Robbiani, D.F., McBride, K., Nussenzweig, M.C., and Casellas, R. (2011). Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. Nat. Immunol. *12*, 62–69.

Yamane, A., Robbiani, D.F., Resch, W., Bothmer, A., Nakahashi, H., Oliveira, T., Rommel, P.C., Brown, E.J., Nussenzweig, A., Nussenzweig, M.C., et al. (2013). RPA accumulation during class switch recombination represents 5′-3′ DNA end resection during S-G2/M phase of the cell cycle. Cell Rep. Published January 02, 2013. http://dx.doi.org/10.1016/j.celrep.2012.12.006.

Zhang, J., MacLennan, I.C., Liu, Y.J., and Lane, P.J. (1988). Is rapid proliferation in B centroblasts linked to somatic mutation in memory B cell clones? Immunol. Lett. *18*, 297–299.

Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S.L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., et al. (2012a). The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature *481*, 157–163.

Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012b). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell *148*, 908–921.

Cell

EXTENDED EXPERIMENTAL PROCEDURES

ChIP-Seq Sample Preparation

At 22 hr poststimulation, cells were treated with 10 mM hydroxyurea for 6 hr, and harvested at 28 hr after stimulation. DNA content was monitored by FACS before proceeding with deep-sequencing analysis. Cells were crosslinked for 10 min at RT with 1% (vol/vol) formaldehyde, followed by quenching with 0.125 M glycine (final concentration). Crosslinked cell samples were then sonicated to obtain DNA fragments 200–500 bp in length using a Covaris S220 sonicator. For each ChIP sample, sonicated chromatin (from 1×10^7 cells) was precleared with 40 μl protein A magnetic beads, followed by incubation overnight at 4°C in radioimmunoprecipitation assay (RIPA) buffer (10 mM Tris, pH 7.6, 1 mM EDTA, 0.1% (wt/vol) SDS, 0.1% (wt/vol) sodium deoxycholate and 1% (vol/vol) Triton X-100) with 40 μl protein A magnetic beads (Invitrogen) and 10 μg of indicated antibody (α-RPA (Calbiochem, RPA34-20); α-BRCA1 (mouse monoclonal raised against mouse BRCA1 (160-300 amino acids); α-SMC5 (Calbiochem, DR1030). Beads were washed twice with RIPA buffer alone, then twice with RIPA buffer plus 0.3 M NaCl, twice with LiCl buffer (0.25 M LiCl, 0.5% (vol/vol) Igepal CA630 and 0.5% (wt/vol) sodium deoxycholate), once with Tris-EDTA buffer (Tris 10 mM, pH 8.0, and 1 mM EDTA) plus 0.2% (vol/vol) Triton X-100, and once with Tris-EDTA buffer alone. ChIP DNA was extracted for 4 hr at 65°C in Tris-EDTA buffer with 0.3% (wt/vol) SDS and proteinase K (1 mg/ml). Anti-γ-H2AX ChIP was performed as described in (Savic et al., 2009) (Millipore, 05-636). Samples were sequenced on a Genome Analyzer IIx after being prepared for microsequencing according to the manufacturer's protocol (Illumina). Each ChIP-seq was performed two times using two separate mice. Each genome-wide analysis represents a single immunoprecipitation analyzed in one lane of the Illumina slide.

Repli-Seq

Cells were cultured as described above in media containing 10 μM BrdU (Sigma). At 22 hr poststimulation, cells were treated with 10 mM HU for 6 hr, and harvested at 28 hr after stimulation. DNA content was monitored by FACS before proceeding with deep-sequencing analysis. Repli-Seq preparation of samples was performed as described (Hansen et al., 2010) using α-BrdU antibodies (GE Healthcare, RPN-202 and BD Biosciences, 347580).

RNA-Seq

Cells were stimulated and treated as above, and total RNA was isolated using TRIzol (Ambion) extraction from 1×10^7 cells following manufacturers' protocol. RNA was washed, purified using RNeasy kit (QIAGEN), and measured for quality using the Agilent RNA 6000 Nano reagents and Bioanalyzer. RNA was then prepared for sequencing using the TruSeq RNA sample prep kit (Illumina). Each RNA-Seq was performed two times using two separate mice.

BACs

Individual BACs to ERFSs were identified using NCBI clone finder, and purchased from BACPAC. BAC probes used in FISH are as follows: MHC2, RP23-208L18; GIMAP, CH29-600D6; SWAP70, RP23-478K21; BACH2, RP24-168B22, IKZF1, CH29-617D21; FOXP1, RP23-352E8; BCL2, RP23-286B14; FRA8E1 (WWOX), RP24-316H14; FRA14A2 (FHIT), RP24-189A10; CNTNAP4, RP24-64E10; SLITRK6, RP24-290G7.

Metaphase Spreads and FISH Analysis

At 22 hr poststimulation, cells from WT, *XRCC2*$^{-/-}$, or *AID*$^{-/-}$ mice were treated with 10 mM HU for 6 hr, and released into fresh media overnight. Cells were then harvested for metaphases and FISH analysis was performed as described in (Callén et al., 2007). For aphidicolin treatment, cells were cultured for 48 hr, then aphidicolin was added to a final concentration of 0.2 μM for approximately 20 hr. For ATR inhibitor treatment, cells were cultured for 48 hr, then ATR inhibitor was added to cells at a final concentration of 1 μM ETP-46464 (Toledo et al., 2011) overnight (14–18 hr).

Quantitative RT-PCR

RNA was isolated from day 3 LPS/IL4-stimulated *SWAP70*$^{-/-}$ and WT littermate B cells using TRIzol (Ambion), and cleaned using RNeasy kit (QIAGEN). cDNA was then made using the Superscript III reverse transcriptase kit (Invitrogen). cDNA was analyzed by RT-PCR using primers to exon 4 of *SWAP70* (p3fwd, 5-GAGGAGGTTGGCAACAAGAA and p1Rev, 5-CCATGCCCTTGCTAAACTGT) on an Applied Biosciences 7900HT. Expression was normalized to β-actin (ACT1, 5-GAGAGGGAAATCGTGCGTGA and ACT2, 5-ACATCTGCTGGAAGGTGG).

ChIP-Seq Reads Processing

Sequence reads of 36 bps were obtained using standard Illumina image-analysis pipeline with default quality filters. The outcome of the pipeline was outputted to the fastq format, trimmed to 25 bps and aligned to the Build 37 assembly of the National Center for Biotechnology Information mouse genome (July 2007; NCBI37/mm9). Bowtie 0.12.8 was used to determine uniquely aligned reads with no mismatches for downstream analysis (Langmead et al., 2009).

ChIP-Seq Sequence Analysis for RPA, BRCA1, SMC5, and γ−H2AX

The unique ChIP-seq reads were mapped into nonoverlapping 200 bp genomic bins. Statistical enrichment of reads' pileup was assessed first relative to those in a Poissonian random background model using the SICER program (Zang et al., 2009) with parameters: e-value, 100, gap size 600 (200) for RPA and γ−H2AX (BRCA1 and SMC5). Then the statistical enrichment of bins passing the first criterion was examined relative to negative control DNA from whole-cell extract based on the threshold of tag density fold increase determined from False Discovery Rate of 1×10^{-5}. Subsequently, the identified enriched 200 bp windows that were less than 5 kbp apart were merged to delineate enriched genomic sites. To generate the custom tracks, the reads within the enriched genomic sites were counted in 20 kbp sliding-windows at 200 bp intervals and adjusted for the library size and presented as reads per million aligned reads (RPM).

ChIP-Seq Sequence Analysis for BrdU

The unique ChIP-seq reads were mapped into nonoverlapping 200 bp windows. Statistical enrichment of reads' pileup was assessed relative to those in a Poisson random background model using the SICER program (Zang et al., 2009) with parameters: e-value, 100, gap size 600. To generate the custom tracks, the reads within the enriched genomic sites were counted in 20 Kbp sliding-windows at 200 bp intervals and adjusted for the library size and presented as reads per million aligned reads (RPM).

RNA-Sequencing

Sequence reads from each cDNA library were mapped onto the Build 37 assembly of the National Center for Biotechnology Information mouse genome data (July 2007; NCBI37/mm9) using TopHat, and outputted to bam format (Trapnell et al., 2009). Bioconductor (Gentleman et al., 2004) packages were used to quantify the expression abundance of RefSeq genes from the aligned reads and calculate the RPKM (reads per kilobase exon model per million mapped reads). Visualization was achieved by the generation of custom tracks for the University of California at Santa Cruz (UCSC) Genome Browser.

Binding Distribution across Selected Set of Coordinates

To show distribution of a given library (query library, indicated on the top) across selected set of coordinates (reference library, indicated on the *x* axis) and their flanking regions, we divided the flanking window surrounding the center of each reference library site (shown bellow each column) into bins of specific size (Figures 1C, 1D, S1B, and S1C). We then counted the number of tags that fall into each bin using filtered aligned reads (those reads that lie inside enriched genomic sites). The outcome matrix represents the tag distribution of the query library across the set of reference coordinates. K-mean method was used to cluster the rows of outcome matrix and generate groups of reference coordinates based on the distribution of the query library. R function 'pheatmap' was used for generating the heatmap after clustering the reference coordinates based on the tag distribution of the query library. A colormap corresponding to number of tags falling into bins was used to show the tag enrichment. For example, the visualization in Figure 1D shows the read distribution of 10 mM HU-treated RPA, BRCA1, and SMC5 libraries (query libraries, on the top) across 10 mM HU-treated RPA-bound sites (reference library, indicated on the *x* axis) and their ±10 kbp flanking regions. The 10 kbp window was divided to 100 bins of 200 bp size.

Correlation of ERFS with Genomic Features

The classes of repeat sequences identified using Repeatmasker and RefSeq gene annotation were downloaded from the UCSC Genome Browser (Figures 1F, S1E, S4D, and S4E). For any given genomic feature, the number of ERFS overlapping the genomic feature of interest was counted. To assess the significance of the observed overlap, a permutation model of ERFS was constructed, that accounts for gaps and the not mappable part of the mouse genome by excluding them from the possible enumerations of the model. For an iteration of the model, a randomized set of ERFS, keeping their lengths and chromosome unchanged, was generated and the count of overlap between the simulated ERFS and the feature of interest was determined. The permutation analysis was repeated to construct the probability model of overlap between the ERFS and the genomic feature.

The permutation analysis was conducted 1×10^3 times for each class of repeat sequences and, the ratio of observed overlap count normalized to the expected overlap count is shown. For RefSeq gene sequences, the permutation analysis was repeated 1×10^5 times and the expected fraction of intergenic and intragenic sequences were plotted and compared to the observed values. Here, the genome was divided into promoters (−5 kbp and +500 bp of TSS), intergenic (>5kbp of TSS), intragenic (+5 kbp of TSS to TES).

Correlation of ERFS with CpG Islands

"CpG islands" were defined as regions of at least 200 bp in length with 60% CG content and a ratio of CpG observed/CpG expected greater than 0.6 (Figure 1G). The "newcpgreport" function of EMBOSS running on NIH biowulf cluster was used to compute the whole-genome CpG islands. The breadth of CpG islands in ERFSs were compared with a permutation model of ERFS that accounts for gaps and not mappable part of the mouse genome by excluding them from the possible enumerations of the model. In an iteration of the model, a randomized set of ERFSs, keeping their lengths and chromosome unchanged, was generated and the breadth of CpG islands in the simulated ERFS was calculated. The permutation analysis was repeated 1×10^5 times to construct the probability model of CpG islands in ERFSs.

Density Plot of RNA-Sequencing Reads

The average read density was calculated in 50 bp bins ±20 kb around the RPA-bound sites' centers (Figure S1G) and ERFS centers (Figure 1H). The average density was then normalized per million aligned reads and the window size, and the loess-smoothed curves (span = 0.1) were plotted.

Correlation of ERFS with Convergent/Divergent Transcripts

A pair of annotated RefSeq genes located on opposite DNA strands was designated as divergent transcripts if their transcription end sites (TES) were within 5 kbp of opposite flanks of an ERFS center, or their intragenic regions overlapped (Figure 1I). A convergent pair of transcripts was defined as two transcripts on opposite DNA strands whose transcription start sites (TSS) were within 5 kbp of opposite flak of an ERFS center. Counts of divergent and/or convergent transcript pairs overlapping with ERFS was calculated and compared with a permutation model repeated 1×10^5 times of simulated ERFS sets.

Definition and Ranking of ERFS Hot Spots

Genomic regions with minimum 10 kbp extent that covered by cluster of ERFS within 300 kbp of one another were defined as ERFS hot spots (Figure 2; Table S1). The average weighted read counts of RPA, BRCA1, and SMC5 of all the ERFS within each hot spot were computed, and their maximum value was assigned as the ranked statistics of hot spot. The ERFS hot spots were ranked based on this score. The aforementioned definitions of convergent and divergent transcript pairs were used to identify the convergent (divergent) open reading frames that are at most 20 kbp apart. A convergent/divergent gene pair was designated as expressed if its both open reading frames had RPKM greater than 1. Genes within 10 kbp flanking sequences of a hot spot were designated as ''Coinciding Genes,'' and the ones with more than three coinciding genes were labels as ''Gene Cluster.'' A hot spot was considered ''Transcriptionally Active'' if the total RPKM of its coinciding genes were greater than 6. Statistically significant AID-dependent genes based on endogenous and overexpression level of AID were obtained from (Chiarle et al., 2011) and (Hakim et al., 2012), respectively. These lists were queried against the ERFS hot spots and the result is reported in Table S1 and Figure 2C. An ERFS hot spot containing AID-dependent genes based on the endogenous expression level of AID was marked as ''AID Target (wild-type).'' An ERFS hot spot containing AID-dependent genes based on the overexpression level of AID in transgenic mice was marked as ''AID Target (AID overexpression).'' An ERFS hot spot was designated as ''AID Target'' (Figure 2C) whenever it contains AID-dependent genes based on either data set.

Overlap of ERFS and Minimal Common Regions of DLBCL

''Minimal common regions'' (MCR) of DLBCL were obtained from (Lenz et al., 2008) and were refined further by considering only the ones with maximum extent of 21 Mbp (Figures 5F and 5G). The ERFS genomic coordinates were overlaid onto Build 36 assembly of the National Center for Biotechnology Information human genome data (March 2006; NCBI36/hg18) using the ''liftOver'' and ''pslMap'' programs. The extent of overlap between the MCR and ERFS is reported, and their significances were compared to permutation models of ERFS and MCR. Each permutation model accounted for gaps and not mappable part of the genome by excluding them from the possible enumerations of the model. The permutation analysis was repeated 1×10^6 times to construct the probability models of overlap between the ERFS and MCRs and vice versa between the MCR and ERFS. The result is shown in Figure 5F.

Using the constructed probability model, the expected overlap of MCR with the permutation model of simulated ERFS was compared to the observed overlap between ERFS and MCR. The percentage of increase in the overlap between the ERFS and MCR relative to the permutation model is reported in Figure 5G.The list of validated CFS was obtained from (Bignell et al., 2010). The overlap between ERFS and MCR was compared to the overlap of CFS with MCR as described below, to normalize for the difference between their genomic size. For each class of fragile sites (ERFS and CFS) with a total number of regions, N, and a total genomic length of L base pairs, the observed overlap with MCR, n, was counted. The expected number of overlaps between the fragile sites and MCR, m, was calculated under a uniform null distribution of the fragile sites across the human genome of total length, G (excluding the gaps and not mappable parts). The enrichment of overlap between each class of fragile sites relative to expected was defined as e = (n-m)/m. The percentage of increase in the overlap between the ERFS and MCR relative to the CFS was defined as (e_MCR – e_CFS)/e_CFS and is shown in Figure 5G.

Overlap of ERFS and Copy-Number Alterations Identified in T-ALL

The copy-number alterations identified by analysis of SNP microarray data of T-ALL were obtained from (Zhang et al., 2012) and further refined by considering only the ones with maximum extent of 21 Mbp (Figure S6A). The ERFS genomic coordinates were overlaid onto Build 36 assembly of the National Center for Biotechnology Information human genome data (March 2006; NCBI36/hg18) using the ''liftOver'' and ''pslMap'' programs. The extent of overlap between the ERFS and copy-number alterations of T-ALL was reported and compared to the overlap of ERFS with copy-number alterations in DLBCL.

Association between ERFS and Putative Cancer Genes with Reported Deletion/Amplification

Known cancer genes were obtained from (Bignell et al., 2010) and their associations with the ERFS were examined (Figure 5H).

Global Reproducibility of ChIP-Seq Data Set

Tag density of ChIP-seq library across 200 kbp nonoverlapping genomic bins were found, and used to plot the correlation of repeat experiments (Figures S1A, S3B, and S3C). R was used to compute the Spearman correlation between the repeats.

Expression at RPA-Bound Genes

A total of 10 mM HU-treated RPA read counts were used to identify RefSeq genes (+5 kbp of TSS to TES) with or without RPA binding (Figure S1H). Genes with a positive RPA read count were denoted as RPA-bound. The box-plot presentation of the expression of RPA-bound and RPA-unbound genes shows the difference between their gene expression distributions. p Value was computed using Wilcoxon rank sum test (wilcox.test in R).

Comparison of G+C Percentage in ERFS to the Whole Mouse Genome

The sequences of Build 37 assembly of the National Center for Biotechnology Information mouse genome (July 2007; NCBI37/mm9) were downloaded (Figure S4B). The whole genome fraction of G and C nucleotides was calculated and compared to the distribution of G and C nucleotides fraction observed in ERFS.

Repli-Chip-Based Replication Timing Profiles

Wavelet-smoothed signal from Repli-chip of CH12 B cell lymphoma and Tc1 $CD4^+$ cell lines were obtained (Stamatoyannopoulos et al., 2012; Figures S1D, S4C, and S5B). The regions of early replication were identified using a zero-crossing algorithm that detects the transition from early to late replicating genomic regions. The early replicating genomic regions were visualized by producing the custom tracks for the UCSC. The coordinates of 803 early replicating genomic regions in CH12 cell line were compared against RPA-bound sites and ERFS.

SUPPLEMENTAL REFERENCES

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. *5*, R80.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Savic, V., Yin, B., Maas, N.L., Bredemeyer, A.L., Carpenter, A.C., Helmink, B.A., Yang-Iott, K.S., Sleckman, B.P., and Bassing, C.H. (2009). Formation of dynamic gamma-H2AX domains along broken DNA strands is distinctly regulated by ATM and MDC1 and dependent upon H2AX densities in chromatin. Mol. Cell *34*, 298–310.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105–1111.

Zang, C., Schones, D.E., Zeng, C., Cui, K., Zhao, K., and Peng, W. (2009). A clustering approach for identification of enriched domains from histone modification ChIP-Seq data. Bioinformatics *25*, 1952–1958.

Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S.L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., et al. (2012). The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature *481*, 157–163.

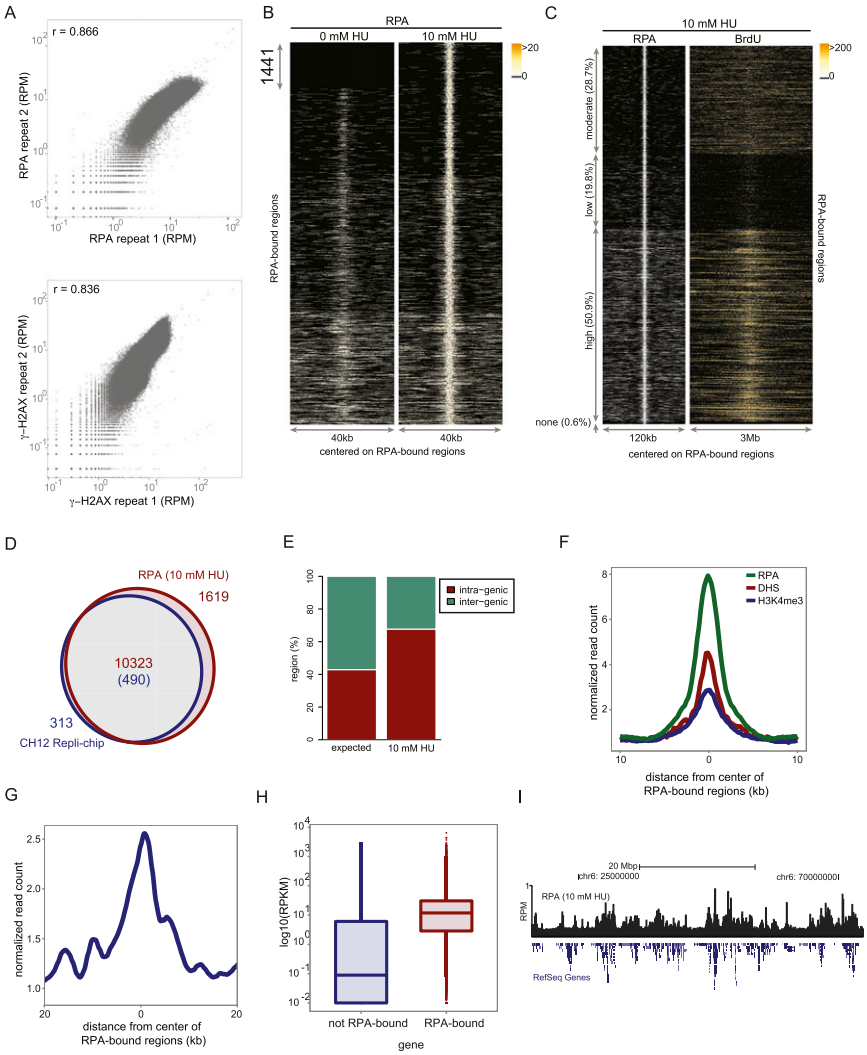


Figure S1. Global Binding of RPA Maps the Early Origins of Replication in HU-Synchronized Cells, Related to Figure 1

(A) Top: global comparison of RPA ChIP-seq experiments between biological duplicates reveals reproducibility of RPA data. Bottom: biological replicates of γ-H2AX ChIP-seq. The Spearman's correlation coefficient between duplicates over 200 kbp sliding windows covering the genome is indicated in the insert.

(B) HU treatment increases the global RPA recruitment to chromatin. For each RPA-bound site in response to HU treatment (y axis), each column depicts the presence of RPA in the untreated (left) and 10 mM HU-treated cells (right) within a window centered on the RPA-bound sites. Color map corresponds to binding intensities where "black" represents no binding.

(C) RPA-bound sites coincide with the early activating replication origins. For each RPA-bound site (y axis), each column depicts the presence of RPA (left) and BrdU incorporation (right) within a window centered on the RPA-bound sites. Color map corresponds to binding intensities where "black" represents no binding. The fraction of each cluster indicating the extent of BrdU incorporation is provided on the left side bar.

(D) RPA-occupied genomic sites significantly overlap with early replicating regions mapped in the CH12 B cell line. Venn diagram depicts the number of shared and unique RPA-bound sites in primary B cells and early replication regions in the CH12 cell line ($p(\text{permutation}) < 1 \times 10^{-5}$) (Stamatoyannopoulos et al., 2012). The total number of regions is indicated for each shared and unique area and color-coded based on the region's title.

(E) RPA-bound sites are enriched in intragenic sequences. Genomic distribution of RPA-bound sites in intragenic (+5 kbp of transcription start site [TSS] to transcription end site [TES]) and intergenic regions is compared to the permutation model ($p(\text{permutation}) < 1 \times 10^{-5}$).

(F) RPA is enriched at DHS and H3K4me3-bound sites. The average normalized read count (loess-smoothed) of RPA, H3K4me3, and DNase-Seq in a 100bp window across a 20 kbp genomic region centered on the RPA-bound sites is plotted.

(G) The line plot represents the average RNA tag count (loess-smoothed) in a genomic window around the center of the RPA-bound sites.

(H) RPA-bound protein coding genes are transcribed more than those lacking RPA binding. Box plots show the first, median and third quantiles of gene expression level in log10 (PRKM) at genes (+5 kbp of TSS to TES) with RPA binding (right) versus genes without RPA binding (left) ($p(\text{Mann-Whitney}) < 1 \times 10^{-16}$).

(I) Characterization of unfiltered RPA ChIP-seq reads. Histogram of all the RPA ChIP-seq reads at chromosome 6 in 10 mM HU-treated B lymphocytes. Gene tracks represent RPA (black) and RefSeq annotated genes (blue), in that order from the top, spanning the entire chromosome 6. The x axis represents the linear sequence of genomic DNA, and the y axis represents the RPM in 200 nucleotide windows (sliding-window smoothed). The genomic scale in mega-bases pairs (Mbp) is indicated above the track.

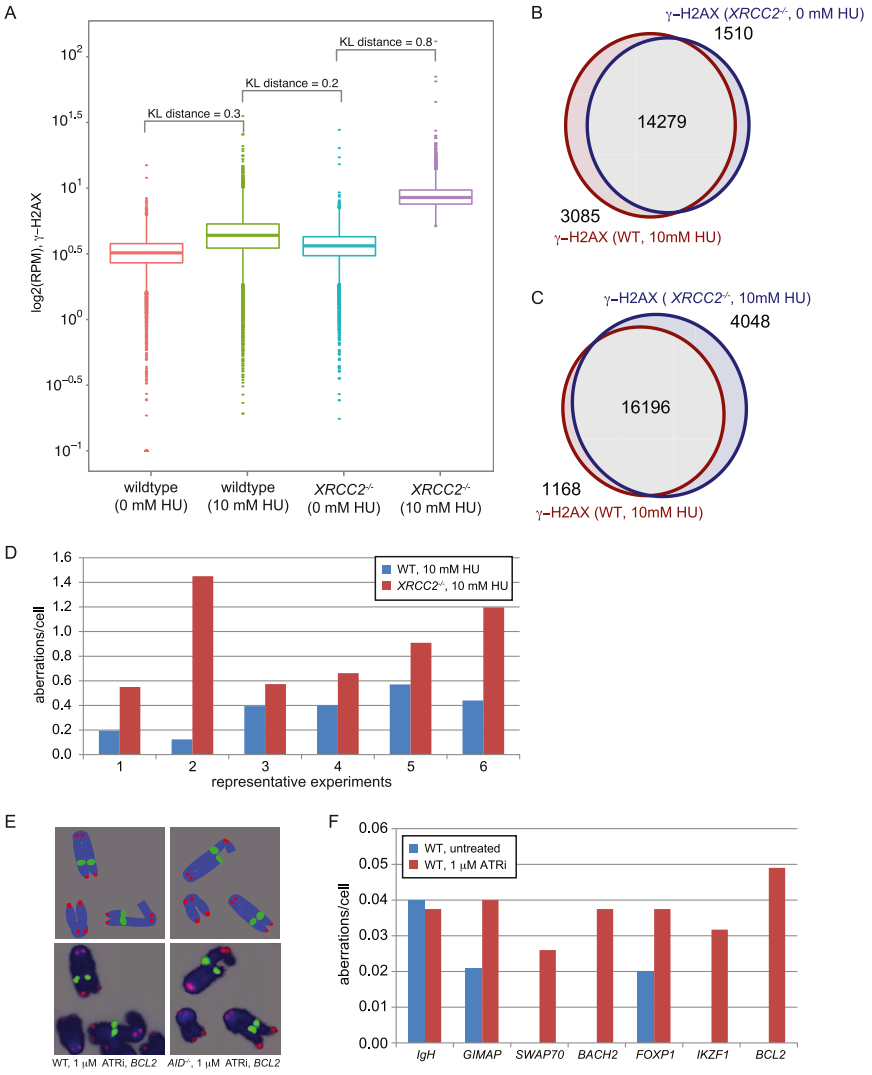


Figure S2. *XRCC2*-Deficient Cells Are Sensitive to Replication Stress, Spontaneously and in Response to HU Treatment, Related to Figure 3

(A) *XRCC2*$^{-/-}$ cells exhibit higher γ-H2AX at the shared loci both spontaneously and in response to HU. Box plots show the first, median and third quantiles of γ-H2AX normalized read count (RPM) in each designated genotype/treatment (x axis) on the shared genomic regions. The Kullback-Leibler divergence between the distributions of γ-H2AX is indicated in the insert.

(B) *XRCC2*$^{-/-}$ cells exhibit accumulation of γ-H2AX under normal growth conditions at similar genomic regions observed in WT 10 mM HU-treated cells. Venn diagram depicts the number of shared and unique γ-H2AX-bound sites in WT HU-treated and *XRCC2*$^{-/-}$ with normal growth conditions.

(C) Significant overlap is observed between the γ-H2AX-occupied sites in WT and *XRCC2*$^{-/-}$ in response to 10 mM HU treatment. Venn diagram depicts the number of shared and unique γ-H2AX-bound sites in WT and *XRCC2*$^{-/-}$.

(D) Quantitation of total DNA aberrations observed in 6 different HU experiments in WT and *XRCC2*$^{-/-}$ cells.

(E) Example of chromosomal aberrations in WT and *AID*$^{-/-}$ cells after treatment with ATRi. Upper: diagram of FISH probes. Lower: representative metaphase showing DNA breaks in the vicinity of the *BCL2* locus in both a WT cell (left) and in an *AID*$^{-/-}$ cell (right) in response to treatment with ATRi. Note that the breakpoint is observed on the centromeric and telomeric side of the BAC in the WT and *AID*$^{-/-}$ cell respectively.

(F) Aberrations per cell at *IgH, GIMAP, SWAP70, BACH2, FOXP1, IKZF1,* and *BCL2* loci in WT cells that were either untreated (blue) or treated with 1 μM ATRi (red) for 24 hr overnight starting 2 days after stimulation with LPS+IL4.

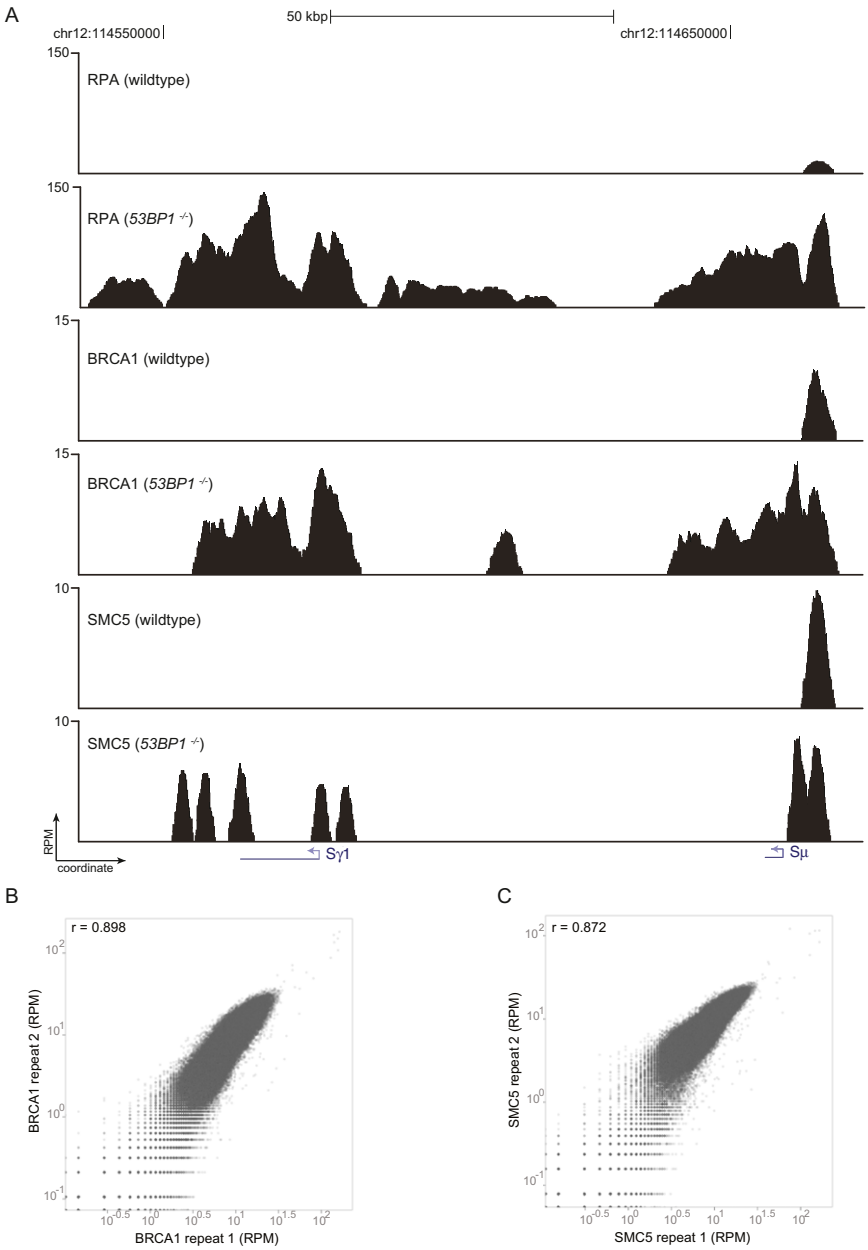


Figure S3. Evaluation of SMC5 and BRCA1 ChIP Antibodies, Related to Figure 1

(A) ChIP-seq read histogram of RPA, BRCA1 and SMC5 at the *IgH* locus in WT (repair proficient) and *53BP1*$^{-/-}$ (repair deficient) cells. Gene tracks represent, from the top, RPA, BRCA1, and SMC5 binding at the region spanning Sμ and Sγ1 switch loci of the IgH locus in WT and *53BP1*$^{-/-}$ cells. The x axis represents the linear sequence of genomic DNA, and the y axis represents the RPM in 200 nucleotide windows (sliding-window smoothed). The genomic scale in kilobase pairs (kbp) is indicated above each track.

(B and C) Global comparison of BRCA1 and SMC5 ChIP-seq experiments between biological duplicates. The Spearman's correlation coefficient between duplicates over 200 kbp sliding windows covering the genome is indicated in the upper left side of the panels.

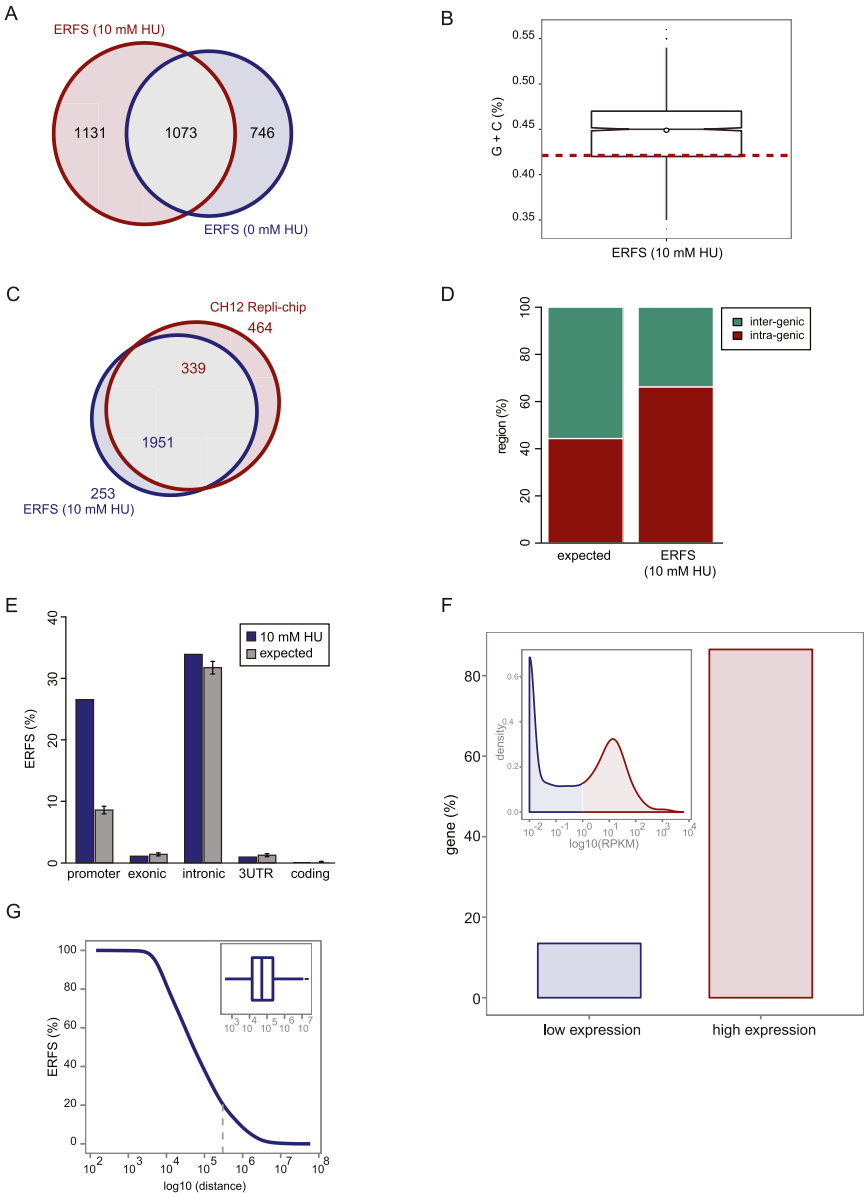


Figure S4. Characterization of RPA-BRCA1-SMC5 triple Colocalized ERFS Regions, Related to Figure 1

(A) Venn diagram depicts the number of shared and unique RPA/SMC5/BRCA1 triple-bound sites in 10 mM HU-treated and untreated conditions.

(B) G+C nucleotides are significantly enriched in ERFSs. Box plot shows the first, median and quantiles of the fraction of G and C nucleotides in the ERFS genomic regions. Dashed line indicates the expected G and C fraction in the whole mouse genome (p(Wilcoxon) < 1×10^{-16}).

(C) Venn diagram depicts the number of shared and unique ERFS in primary B cells and early replication regions identified in the CH12 B cell line (Stamatoyannopoulos et al., 2012). The total number of regions is indicated for each shared and unique area and color-coded based on the regions title.

(D) ERFS regions are enriched at intragenic sequences. Genomic distribution of ERFS sequences in intragenic (+5 kbp of TSS to TES) and intergenic regions is compared to the permutation model (p(permutation) < 1×10^{-3}).

(E) ERFS regions are enriched at promoter sequences. Genomic distribution of ERFS sequences in translated and untranslated sequences of protein coding genes is compared to the permutation model. Blue bars show the fraction of ERFS sequences. Gray bars show the expected fraction of each genomic sequence ± standard deviations based on the permutation model.

(F) Genes located within ERFS are highly transcribed. The top-left insert shows the probability distribution function of gene expression levels in 10 mM HU-treated cells. The set threshold (RPKM = 1) categorizes genes as having either high or low expression level. Bar graphs show the distribution of genes within ERFSs designated as either high or low transcription (p(binomial) < 1×10^{-16}).

(G) ERFS are structured into clusters. Plot shows the empirical cumulative distribution of distances between ERFS. Gray line depicts the 80-percentile of the distribution. Box plot indicating the first, median and third quantiles of the distribution is shown in the top-right insert.

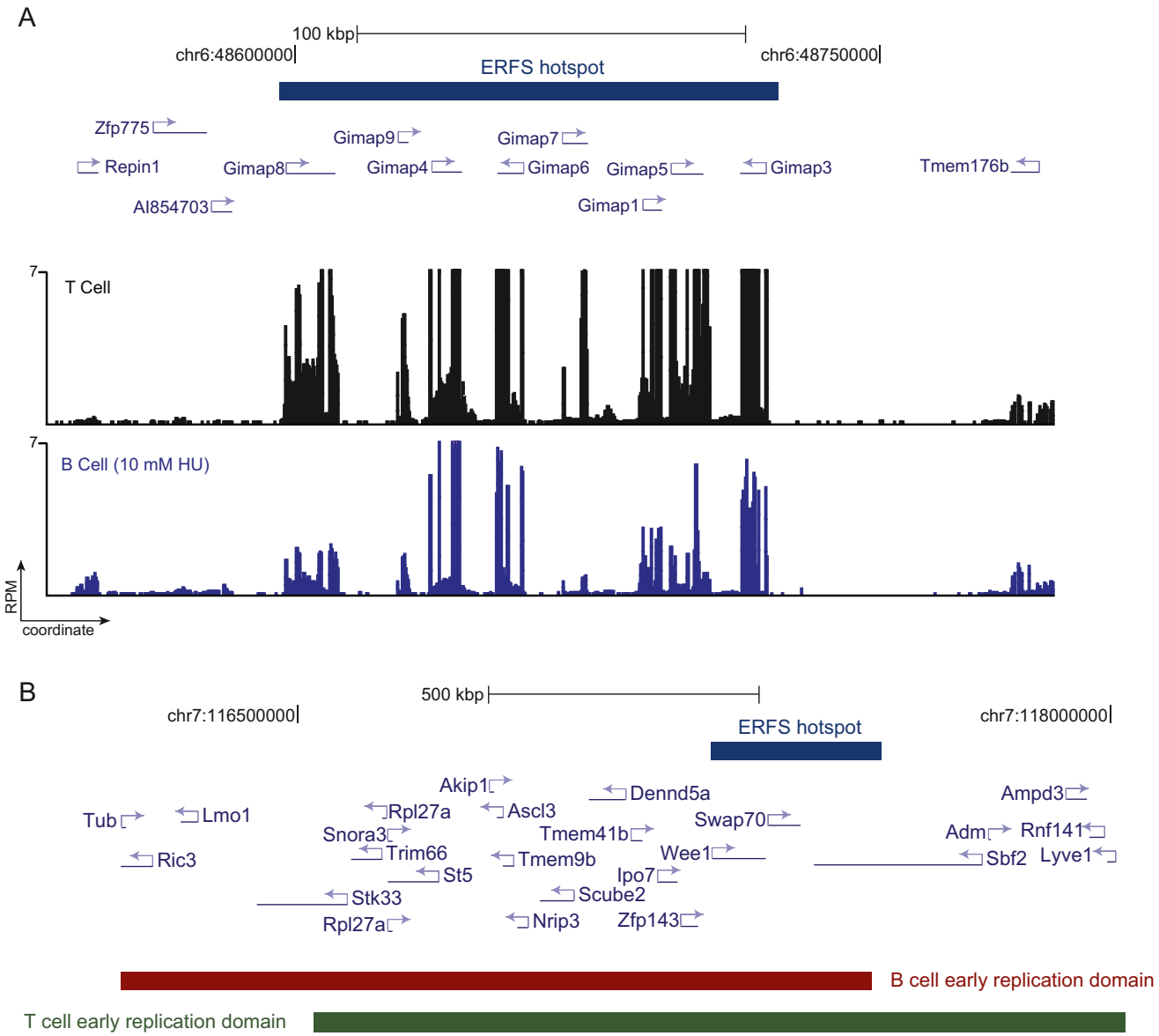


Figure S5. Expression of *GIMAP* and Early Replication near *SWAP70* in B and T Cells, Related to Figure 4

(A) Gene tracks represent, from the top, ERFS demarcation, and transcription measured by RNA-Seq in T and B cells at the region spanning around *GIMAP* locus. The *x* axis represents the linear sequence of genomic DNA, and the *y* axis represents the RPM.

(B) The *SWAP70* locus replicates early in both B and T cells. Gene tracks represent, from the top, ERFS demarcation, open reading frames, and early replicating regions mapped in the CH12 B cell and the Tc1 CD4$^+$ T cell lines (Stamatoyannopoulos et al., 2012). The *x* axis represents the linear sequence of genomic DNA.

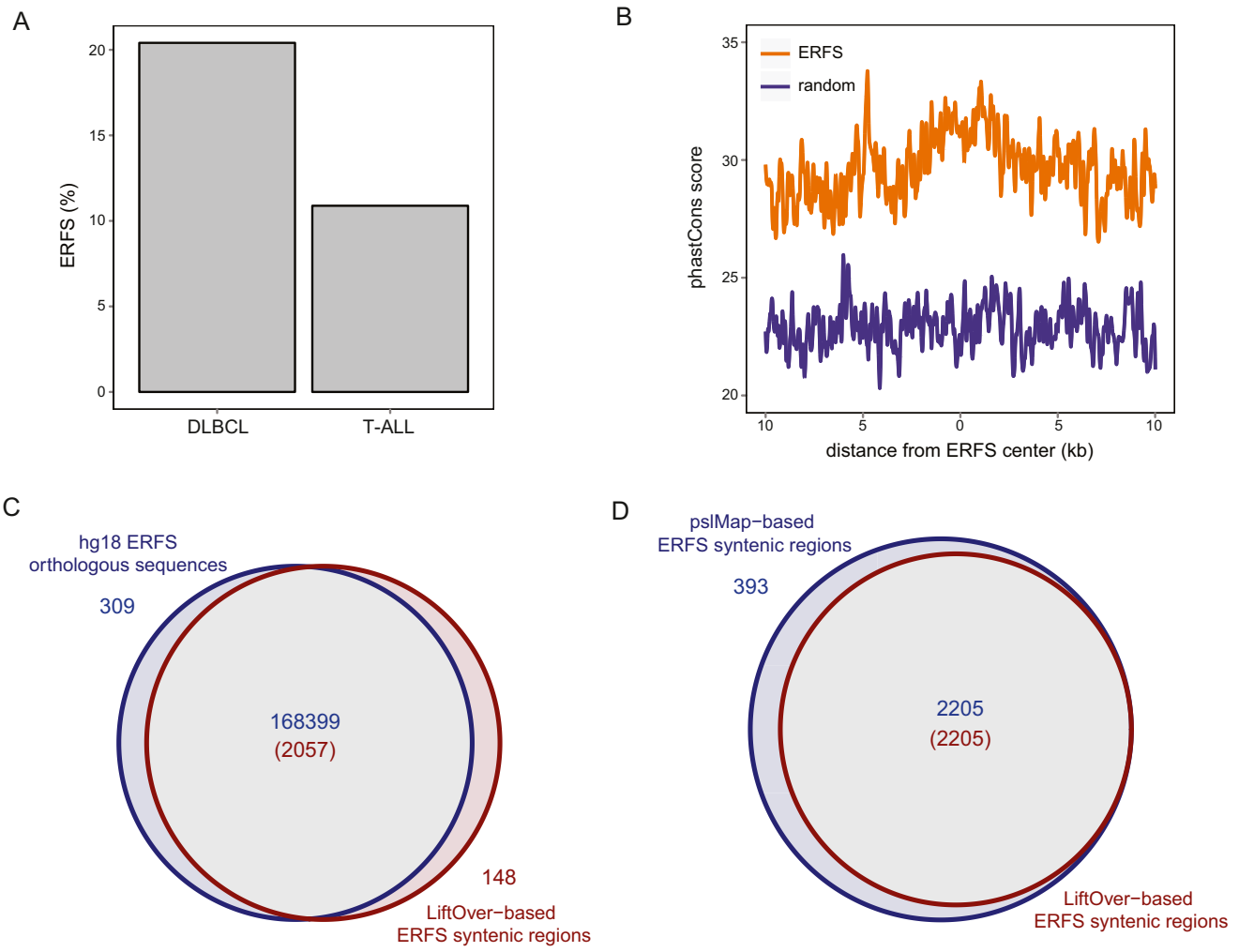


Figure S6. Characterization of ERFS Synteny Map, Related to Figure 5

(A) ERFS correlation with copy-number alterations is cancer type-specific. The extent of overlap between the ERFSs and regions of copy-number alterations detected in DLBCL and T-ALL is assessed, and depicted in the left and right bar-graphs, respectively.

(B) Enrichment of evolutionary conserved sequences based on phastCons score around the center of ERFS (red) compared to random regions of the genome (blue) (±10 kbp the center or ERFS region) (p(Kolmogorov-Smirnov) < 2×10^{-16}).

(C) The Venn diagram shows the overlap of LiftOver-based syntenic ERFS regions and the identified orthologous ERFS sequences in the human genome NCBI Build 36/hg18 (euarchontoglires mammals multiple alignment) (Zhang et al., 2012). The total number of regions is indicated for each shared and unique area and color-coded based on the region's title. The areas are not proportionally rendered.

(D) The Venn diagram shows the overlap of two synteny maps of ERFS regions in NCBI Build 36/hg18 constructed using LiftOver and pslMap methods. The total number of regions is indicated for each shared and unique area and color-coded based on the region's title.