Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing

Nicola Crosetto1,2,11,12, Abhishek Mitra3,12, Maria Joao Silva4, Magda Bienko1,2,11, Norbert Dojer3,11, Qi Wang5,6, Elif Karaca5,6, Roberto Chiarle5–7, Magdalena Skrzypczak8, Krzysztof Ginalska8, Philippe Pasero4, Maga Rowicka3,9,10 & Ivan Dikic1,2

We present a genome-wide approach to map DNA double-strand breaks (DSBs) at nucleotide resolution by a method we termed BLESS (direct in situ breaks labeling, enrichment on streptavidin and next-generation sequencing). We validated and tested BLESS using human and mouse cells and different DSBs-inducing agents and sequencing platforms. BLESS was able to detect telomere ends, Sce endonuclease–induced DSBs and complex genome-wide DSB landscapes. As a proof of principle, we characterized the genomic landscape of sensitivity to replication stress in human cells, and we identified >2,000 nonuniformly distributed aphidicolin-sensitive regions (ASRs) overrepresented in genes and enriched in satellite repeats. ASRs were also enriched in regions rearranged in human cancers, with many cancer-associated genes exhibiting high sensitivity to replication stress. Our method is suitable for genome-wide mapping of DSBs in various cells and experimental conditions, with a specificity and resolution unachievable by current techniques.

DNA DSBs can be caused by exogenous or endogenous physical or chemical agents, and they appear during apoptosis, meiotic crossing-over and gene rearrangements1–2. Replication-fork stalling and collapse also causes DSBs and is considered the major endogenous source of breaks in cycling cells3,4. Unresolved DSBs pose a serious threat to genomic stability, potentially leading to the formation of oncogenic mutations, including translocations, deletions and amplifications2.

Although past research has provided extensive knowledge on mechanisms of DSB sensing and repair, the genome-wide landscape of DSBs in different cell types and conditions remains largely unknown, mainly because of the lack of methods to map DSBs with high specificity and resolution throughout the genome. Chromatin immunoprecipitation (ChiP) coupled to microarray (ChiP-on-chip) or next-generation sequencing (ChiP-seq) has been applied to map DSBs5–7. However, the fact that in ChiP-based methods DSBs are not directly labeled in situ, but rather detected indirectly using antibodies targeting specific proteins that bind to DSBs, represents a considerable source of bias. The phosphorylated histone variant H2A.X (γH2A.X) is typically used as a marker of DSBs, but it can also mark single-strand breaks and the inactive X chromosome8–10. Moreover, γH2A.X ChiP signals can spread tens of kilobases away from a single DSB11,12, making it difficult to map the exact position of a DSB. Alternatively, the recruitment of replication protein A (RPA) has been used to map DNA damage, but RPA accumulation is partially blocked by 53BP1 (refs. 13,14), thereby limiting the reliable use of RPA to cells lacking 53BP1 (ref. 15). Other approaches have used capture or direct labeling of ssDNA followed by microarray analysis, assuming that ssDNA is a good proxy for DSBs16–18. A bias in these methods, however, is that any ssDNA not converted to DSBs (for example, during replication) will yield false positive results. Recently, a method based on breaks labeling with the terminal deoxynucleotidyltransferase enzyme has been used to detect DSBs at defined locations in vitro and in purified genomic DNA from Saccharomyces cerevisiae19. This method, however, has not been implemented on a genomic scale, is not in situ and does not allow labeling of DSBs with specific barcode sequences, which would be extremely helpful in filtering next-generation sequencing data. Therefore, genome-wide methods with higher resolution and specificity are needed to gain insights into the biology of DNA DSBs in different cell types and conditions. Here we present a comprehensive experimental and computational approach to directly map DSBs genome-wide, BLESS, based on direct in situ breaks labeling, enrichment on streptavidin and next-generation sequencing.

1Institute of Biochemistry II, Goethe University Medical School, Frankfurt, Germany. 2Buchmann Institute for Molecular Life Sciences, Goethe University Medical School, Frankfurt, Germany. 3Institute for Translational Sciences, University of Texas Medical Branch at Galveston (UTMB), Galveston, Texas, USA. 4Institut de Génétique Humaine (IGH), Centre National de la Recherche Scientifique Unité Propre de Recherche 1142, Montpellier, France. 5Department of Pathology, Children’s Hospital, Boston, Massachusetts, USA. 6Harvard Medical School, Boston, Massachusetts, USA. 7Department of Molecular Biotechnology and Health Sciences, University of Torino, Torino, Italy. 8Laboratory of Bioinformatics and Systems Biology, Centre of New Technologies, University of Warsaw, Warsaw, Poland. 9Department of Biochemistry and Molecular Biology, UTMB, Galveston, Texas, USA. 10Sealy Center for Molecular Medicine, UTMB, Galveston, Texas, USA. 11Present addresses: Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA (N.C. and M.B.); Institute of Informatics, University of Warsaw, Warsaw, Poland (N.D.; on leave). 12These authors contributed equally to this work. Correspondence should be addressed to M.R. (maga.rowicka@utmb.edu; computational analyses) or to N.C. (crosetto@mit.edu) or I.D. (ivan.dikic@biochem2.de; all other correspondence).

Received 10 October 2012; accepted 12 February 2013; published online 17 March 2013; doi:10.1038/nmeth.2408
RESULTS

BLESS workflow

Direct in situ labeling prevents the labeling of DSBs artificially formed during genomic DNA (gDNA) extraction, thus minimizing the risk of false positives. After fixation, performed to stabilize chromatin and prevent artificial DSBs, we lysed the cells and briefly incubated them with proteinase K to purify intact nuclei from both human cell lines and mouse tissues (Supplementary Fig. 1a). We then blunt DSBs, 5′-phosphorylated them and ligated them to a biotinylated linker (proximal) using the highly specific T4 ligase enzyme, which can ligate double-strand but not single-strand breaks. The linker forms a hairpin-like structure and can thus be ligated to either a blunt DSB or an identical linker molecule, but it cannot form concatemers. The ligatable end of the linker consists of a barcode sequence marking the exact position of ligation site (orange and brown, respectively). The proximal linker is biotinylated (orange oval). (c) Proportion of fragments with proximal and distal barcodes. Mean ± s.d. is shown for n = 9 independent single-end (SE) Illumina sequencing experiments (red). Individual percentages are shown for n = 2 paired-end (PE) sequencing experiments (orange and yellow).

BLESS implementation and validation

We performed pilot experiments in HeLa cells and mouse B lymphocytes, which were followed by Sanger sequencing and next-generation sequencing on the Roche 454 platform. We performed various controls to exclude substantial false positive labeling due to incomplete washout of proximal linkers, nonspecific binding of gDNA to streptavidin beads or mispriming (Supplementary Fig. 1b–d). In two biological replicates, 96.6% and 99.9% of barcoded reads contained both proximal and distal barcodes (Supplementary Table 1). As a first proof of specificity, we searched sequences derived from activated mouse B lymphocytes for reads mapping in the immunoglobulin heavy chain locus. Upon B-lymphocyte activation, DSBs are formed in the immunoglobulin heavy chain donor Sμ region and the downstream acceptor S region, enabling antigen class switching 20. Accordingly, the density of correctly barcoded reads within these regions was significantly higher than the average read density in the genome (twofold enrichment, P = 0.02, hypergeometric test), even with the relatively modest throughput achievable with the Roche 454 platform.

To increase data throughput, we performed deeper sequencing of BLESS samples using the Illumina GAII and HiSeq 2000 platforms (Supplementary Table 1). All sequencing data, including Sanger and Roche 454 sequences, can be accessed at http://www.breakome.eu/. In single-end sequencing experiments, the proportions of proximal and distal barcodes among barcoded reads were similar (proximal, 53.9% ± 6.2%; distal, 46.1% ± 4.8%; mean ± s.d., n = 9). Paired-end sequencing of two biological replicates showed that 99.4% and 99.2% of BLESS barcoded fragments contained both proximal and distal barcodes (Fig. 1c and Supplementary Table 1). This result demonstrates that the false positive DSB labeling rate in BLESS is lower than 1%. We initially deep-sequenced HeLa cells, a model system for which a large amount of genome-wide data are available and in which telomeric ends have been well characterized 21. During BLESS, the 3′ G overhang of unprotected telomeres, which resembles a DSB repair intermediate 22, is trimmed down to the first nucleotide of the complementary C strand, where the biotinylated linker is ligated (Supplementary Fig. 2a). Therefore, we expected accessible telomeric ends to be detected by BLESS. Accordingly, we retrieved telomeric reads derived from the C strand, with CTAACC being the most frequent (73%) C-strand end, as previously reported 21 (Supplementary Fig. 2b). We also deep-sequenced U2OS cells carrying an I-SceI transgenic cassette after transfecting them with hemagglutinin-tagged I-SceI to induce a single DSB per cell within this cassette (Supplementary Fig. 2c). The density of barcoded reads inside the I-SceI cassette was almost 13,000-fold higher than the average read density in the genome (P = 6 × 10−300; see Online Methods for derivation of P), further validating our labeling method. These results demonstrate that direct in situ labeling of DSBs followed by next-generation sequencing is an effective strategy to identify DSBs at various genomic locations.

Genomic landscape of sensitivity to replication stress

Deep-sequencing experiments revealed a landscape (‘breakome’) of DSBs sparse throughout the genome. Cells grown in culture carry a non-negligible amount of DNA breaks caused by a combination of replication stress, physiological apoptosis and damage induced by reactive oxygen species 12. Indeed, even in the absence of any exogenous treatment, HeLa cells carried a substantial burden of γH2A.X foci (3.8 ± 6.7 foci per nucleus (mean ± s.d.), n = 300) (Supplementary Fig. 3a,b). To single out breaks caused by replication stress, which are believed to be a major source of genome instability 3, we exposed HeLa cells to a dose of aphidicolin, a DNA polymerase inhibitor, that induces replication-fork

Figure 1 | BLESS workflow and specificity. (a) DSBs are ligated in situ to a proximal linker (red arch) covalently linked to biotin (orange oval) (1), genomic DNA is extracted and fragmented (2) and labeled fragments are captured on streptavidin beads (gray ovals) (3). A distal linker (cyan arch) is then ligated to the free extremity of captured fragments (4), and fragments are released by linker digestion with I-SceI (5). Released fragments are amplified by PCR using linker-specific primers (6) and and fragments are released by linker digestion with I-SceI (5). Released fragments are amplified by PCR using linker-specific primers (6) and and fragments are released by linker digestion with I-SceI (5).
stalling without arresting progression in S phase\textsuperscript{23}. This treatment resulted in significant accumulation of breaks ($P = 10^{-34}$, Kolmogorov-Smirnov test) and increased the amount of labeled DSBs captured by BLESS (Supplementary Fig. 3a–d).

We analyzed aphidicolin sensitivity at various resolutions by comparing deep-sequencing data from replicates of samples that either were or were not treated with aphidicolin (Supplementary Fig. 4a). We compared read numbers within genomic windows with constant mappable length in treated and untreated samples, and we calculated enrichment $P$ values based on the hypergeometric distribution. We computed final $Q$ values based on the Benjamini-Hochberg approach for multiple hypothesis testing\textsuperscript{24} (Online Methods). We mapped ASRs at a resolution of 48 kilobases without and with correction for copy-number variation effects due to karyotype and aphidicolin treatment (Supplementary Fig. 4b), which yielded 2,307 and 2,429 significantly correlated ASRs, respectively ($P = 10^{-323}$, hypergeometric test). To calculate the false discovery rate (FDR) related to our approach, we analyzed reads that were mapped to the Y chromosome through sequencing errors (in HeLa cells, no reads from the Y chromosome are expected). At a resolution of 48 mappable kilobases, the calculated FDR was 0.3%. The full list of ASRs and $Q$ values is available at http://www.breakome.eu/.

ASRs were nonuniformly distributed along the genome, with an average of 3\% of 48 mappable kilobase regions per chromosome being sensitive to aphidicolin, except for those on chromosomes 5 and 7, where the proportion was significantly higher (5\%; $P = 10^{-5}$, hypergeometric test). As a comparison, we applied BLESS to HeLa cells treated with neocarzinostatin, a DSB-inducing drug presumed to yield a more random pattern of breaks. Neocarzinostatin-sensitive regions were significantly more uniformly spread along the genome (distance between consecutive sensitive and aphidicolin treatment: 0.54 ± 1.81 megabases (Mb) and 0.21 ± 0.77 Mb for aphidicolin and neocarzinostatin, respectively, median ± s.d.; $P = 10^{-118}$, Kolmogorov-Smirnov test) (Fig. 2).

To validate our findings, we compared several among the most statistically significant ASRs with regions displaying no appreciable aphidicolin effect using γH2AX ChIP. We observed a strong concordance between the aphidicolin effects measured in targeted regions and BLESS results (Fig. 3). Notably, the genomic locations of ASRs mapped in different experimental replicates were significantly correlated, thus demonstrating the reproducibility of our method (Supplementary Table 2).

Characterization of aphidicolin-sensitive regions

It has been suggested that repetitive DNA sequences may favor fork stalling and collapse upon replication stress, causing DSBs to appear more frequently at certain genomic regions\textsuperscript{3,4}. In particular, repeats prone to form hairpin-like secondary structures might cause the collapse of slowly moving replication forks by directly hindering their progression. To investigate the association between repeats and aphidicolin sensitivity, we applied RepeatMasker\textsuperscript{25} to compute the abundance of various DNA repeat families inside ASRs as compared to the rest of the genome (Online Methods). We detected a reproducible strong and significant enrichment in satellites ($P = 5 \times 10^{-137}$) in particular of alpha-type repeats ($P = 8 \times 10^{-198}$, see Online Methods for derivation of $P$ values), a class of repeats that forms hairpin-like secondary structures and is abundant in pericentromeric and centromeric regions (Fig. 4a, Supplementary Fig. 5 and Supplementary Table 3). Accordingly, ASRs mapped at high resolution (250 mappable nucleotides) were concentrated in pericentromeric and centromeric regions (Supplementary Table 4). Another class of repeats, AT dinucleotides, has been associated with a particular group of genomic regions sensitive to replication stress induced by aphidicolin: common fragile sites\textsuperscript{26,27}. Although many common fragile sites were scored as sensitive to aphidicolin with our approach, AT repeats were significantly depleted in ASRs ($P = 10^{-16}$, hypergeometric test).

During replication stress, slowly moving replication forks will have a higher chance of colliding with transcriptional forks\textsuperscript{28}, resulting in accumulation of DSBs that may be detected by BLESS. Accordingly, upon aphidicolin treatment, we detected a prominent enrichment of DSBs in transcribed regions with the highest...
enrichment in coding regions \( (P = 10^{-10}, \) hypergeometric test). We next analyzed genes and ranked them according to the computed probability of developing DSBs anywhere along their length. The top 20% aphidicolin-sensitive genes showed significant enrichment of many Gene Ontology terms, particularly those related to cell death \( (P = 10^{-3}, \) hypergeometric test). Gene sensitivity to aphidicolin was significantly associated with gene length \( (P = 10^{-18}, \) see Online Methods for derivation of \( P \)), in line with the observation that the probability of collisions between replication forks in active genes seems to increase with gene length \( ^{28, 29} \) (Supplementary Fig. 6).

Finally, we investigated whether ASRs mapped by BLESS are also associated with genomic regions or genes frequently rearranged in human cancers. Replication stress–driven genomic instability has been observed in many tumors, where it is thought to be an important cause of cancer-related genetic rearrangements \(^{30, 31} \). We used data from a cohort of over 2,700 human cancers \(^{32} \) and found a modest but significant enrichment of regions displaying amplifications or deletions inside ASRs as compared to the rest of the genome \( (P = 0.005, \) derived as described in Online Methods) (Fig. 4b). We next compared aphidicolin-sensitive genes with the Cancer Gene Census \(^{33, 34} \), a collection of over 400 well-annotated cancer genes, the majority of which are involved in translocations. Cancer genes were more likely than noncancer genes to overlap with ASRs \( (P = 0.04, \) hypergeometric test), and the fraction of genes with a 5′ end in a 2-Mb vicinity of the center of a 48-mappable-kilobase ASR, and containing that ASR center inside, was higher for cancer genes than others \( (P = 0.02, \) hypergeometric test) (Fig. 4c).

Among most aphidicolin-sensitive genes, cancer genes were overrepresented \( (P = 0.04, \) hypergeometric test), including prominent oncogenes such as \( EGFR, MET, ABL1 \) and \( MLL \), which are typically mutated by translocation or amplification (Fig. 4d). The full list of genes, ranked by aphidicolin sensitivity, and Gene Ontology analysis results are available at http://www.breakome.eu/.

**DISCUSSION**

DNA DSBs represent a major threat to genomic stability, and understanding the sensitivity of the genome to various DNA insults will be instrumental to the implementation of effective preventive and treatment strategies. Replication errors and reactive oxygen species generated as byproducts of metabolism have been estimated to cause breaks at a frequency as high as 50 DSBs per cell per day \(^{12} \). In spite of this pervasive threat, our knowledge on how the genome breaks in response to various insults and our technology for reliably detecting DSBs are still in their infancy. Unlike the empirical background-subtraction procedures used in ChIP-based methods to account for nonspecific binding, direct DSB labeling with the BLESS method ensures high specificity of break detection, which can then be unambiguously identified by the presence of barcode sequences. Another important advantage of BLESS over ChIP-based methods is the ability to directly mark DSBs at nucleotide resolution *in situ*; the latter relies on proxies such as γH2AX, which can be found tens of kilobases away from the actual original DSB \(^{11} \). It should be noted that, at least for DSBs repaired by homologous recombination, labeling can occur away from the initial breakpoint because of 5′ end resection. This property can be exploited by inducing DSBs at known genomic positions to obtain a zoomed-in view into the kinetics of DSB repair *in vivo* and how they are influenced by the genomic context.

Our method is general and organism independent, providing genome-wide maps of DSBs for multiple cell types and conditions. Our computational methods and software tools allow users to obtain and analyze high-confidence genome-wide DSB maps and to account for copy-number variation effects attributable to the karyotype of cells analyzed and/or the effects of the treatment used to induce DSBs. Our results demonstrate that hypothesis-driven feature analysis of genomic regions identified by BLESS can help explore the basis of genomic instability at a genome-wide level. In the future, our method could be combined with ultradepth sequencing of selected regions enriched by, for example, exome capture \(^{35} \) or reduced representation sequencing \(^{36} \), thus providing a high-definition picture of the sensitivity of specific regions to DSB-inducing agents. Finally, the design principle of BLESS could also be exploited for *in situ* DSB labeling and visualization by super-resolution microscopy. The labeling method and the computational approaches described here represent a valuable

---

**Figure 4** | Biological characterization of ASRs. (a) Satellite repeats significantly enriched within 48-mappable-kilobase ASRs in comparison to the rest of the genome. All, all satellite repeat classes; Alpha, centromeric satellite repeats; TAR1, telomere-associated repeat 1. Repeat names follow the nomenclature in RepeatMasker \(^{25} \). Bars show enrichments calculated on the basis of \( n = 4 \) pooled samples. Diamonds show enrichments calculated on the basis of \( n = 2 \) pooled samples. (b) Significant enrichment of cancer-associated somatic copy-number alterations. All, all alterations; Amp, amplifications; Del, deletions. Dashed lines represent average genome-wide enrichment. (c) Percentage of cancer (red) and noncancer (blue) genes containing the center of a 48-mappable-kilobase ASR within 2 Mb downstream of the 5′ end. (d) Ranking of aphidicolin-sensitive cancer-associated genes by decreasing sensitivity, expressed as a percentage of the most sensitive gene on the left.

**Table 1** | Biological characterization of ASRs.

| Distance from 5′ end (kb) | Enrichment inside ASRs | Enrichment inside AllAmpDel | Enrichment inside AllAmpDelAsr |
|--------------------------|------------------------|-----------------------------|-------------------------------|
| 0–20                     | 0.5                    | 0.005                        | 0.04                          |
| 20–100                   | 1.0                    | 0.04                         | 0.08                          |
| 100–200                  | 1.5                    | 0.1                          | 0.16                          |

**Figure 4a** | Enrichment inside ASRs.

**Figure 4b** | Significant enrichment of cancer-associated somatic copy-number alterations.

**Figure 4c** | Percentage of cancer (red) and noncancer (blue) genes containing the center of a 48-mappable-kilobase ASR within 2 Mb downstream of the 5′ end.

**Figure 4d** | Ranking of aphidicolin-sensitive cancer-associated genes by decreasing sensitivity, expressed as a percentage of the most sensitive gene on the left.
resource for the DNA damage research community, providing tools to map and analyze breakomes in a variety of organisms and conditions with a precision and resolution currently unattainable with other methodologies.

**METHODS**

Methods and any associated references are available in the online version of the paper.

**Accession codes.** NCBI Sequence Read Archive: SRP018506.

**Note:** Supplementary information is available in the online version of the paper.

**ACKNOWLEDGMENTS**

We acknowledge Y. Shiloh (Tel Aviv University) and A.J. Pierce (University of Kentucky) for kindly providing U2OS-DRH-1 cells and I-Sce plasmids. We are grateful to T. Włodarski, A.R. Lehmann, G. Fudenberg and A. Kudlicki for insightful discussions, critical reading of the manuscript and help with data analysis. This work was supported by grants from Deutsche Forschungsgemeinschaft, the Cluster of Excellence “Macromolecular Complexes” of the Goethe University Frankfurt (EXC115), the LOEWE-funded Oncogenic Signaling Frankfurt network, the LOEWE Gene and Cell Therapy Center and the European Research Council (ERC) under the European Union’s Seventh Framework Programme (FP7/2007-2013) and ERC grant agreement 250241-LineUb to I.D.; from the Associazione Italiana per la Ricerca sul Cancro (AIRC) and the International Association for Cancer Research (AICR) and grant FP7 ERC-2009-StG (proposal 242965—“Lunely”) to R.C.; from the Foundation for Polish Science (TEAM), Polish National Science Centre (2011/02/A/NN2/00014) and European Regional Development Fund under the Innovative Economy Programme (POIG.02.02.00-14-024/08-00) to K.G.; from Ligue contre le Cancer (équipe labellisée), Agence Nationale de la Recherche (RepliCare and Institut National du Cancer to P.P.; and by grant UL1TR000071 ITS “Novel Programme (POIG.02.02.00-14-024/08) to K.G.; from Ligue contre le Cancer (équipe labellisée), Agence Nationale de la Recherche (RepliCare and Institut National du Cancer to P.P.; and by grant UL1TR000071 ITS “Novel Methods” from the National Center for Research Resources, US National Institutes of Health, to M.R. M.B. is a recipient of a Human Frontier Science Program Long-Term Fellowship.

**AUTHOR CONTRIBUTIONS**

N.C. and I.D. conceived and developed BLESS, coordinated the project and wrote the manuscript. A.M. developed all necessary code and analyzed Illumina data. M.J.S. and P.P. performed ChIP experiments and analysis. M.B. performed microscopy experiments and prepared figures. Q.W., E.K. and R.C. performed Roche 454 experiments and analyzed the data. N.D. contributed to statistical data analysis. M.S. and K.G. performed paired-end Illumina sequencing. M.R. conceived procedures for computational analysis, supervised the analysis and coordinated the project.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Paigen, K. & Petkov, P. Mammalian recombination hot spots: properties, control and evolution. Nat. Rev. Genet. 11, 221–233 (2010).
2. Jackson, S.P. & Bartek, J. The DNA-damage response in human biology and disease. Nature 461, 1071–1078 (2009).
3. Branzei, D. & Folaní, M. Maintaining genome stability at the replication fork. Nat. Rev. Mol. Cell Biol. 11, 208–219 (2010).
4. Branzei, D. & Folaní, M. The DNA damage response during DNA replication. Curr. Opin. Cell Biol. 17, 566–575 (2005).
5. Szilárd, R.K. et al. Systematic identification of fragile sites via genome-wide location analysis of gamma-H2AX. Nat. Struct. Mol. Biol. 17, 299–305 (2010).
6. Harrigan, J.A. et al. Replication stress induces 53BP1-containing OIP domains in G1 cells. J. Cell Biol. 193, 97–108 (2011).
7. Seo, J. et al. Genome-wide profiles of H2AX and -H2AX differentiate endogenous and exogenous DNA damage hotspots in human cells. Nucleic Acids Res. 40, 5965–5974 (2012).
8. Marti, T.M., Hefner, E., Feene, L., Natale, V. & Cleaver, J.E. H2AX phosphorylation within the G1 phase after UV irradiation depends on nucleotide excision repair and not DNA double-strand breaks. Proc. Natl. Acad. Sci. USA 103, 9891–9896 (2006).
9. Tuduri, S. et al. Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription. Nat. Cell Biol. 11, 1315–1324 (2009).
10. Chadwick, B.P. & Lane, T.F. BRCA1 associates with the inactive X chromosome in late S-phase, coupled with transient H2AX phosphorylation. Chromosoma 114, 432–439 (2005).
11. Iacovoni, J.S. et al. High-resolution profiling of γH2AX around DNA double strand breaks in the mammalian genome. EMBO J. 29, 1446–1457 (2010).
12. Bonner, W.M. et al. γH2AX and cancer. Nat. Rev. Cancer 8, 957–967 (2008).
13. Bunting, S.F. et al. 53BP1 inhibits homologous recombination in BRCA1-deficient cells by blocking resection of DNA breaks. Cell 141, 243–254 (2010).
14. Bothmer, A. et al. 53BP1 regulates DNA resection and the choice between classical and alternative end joining during class switch recombination. J. Exp. Med. 207, 855–865 (2010).
15. Hakim, O. et al. DNA damage defines sites of recurrent chromosomal translocations in B lymphocytes. Nature 484, 69–74 (2012).
16. Blitzblau, H.G., Bell, G.W., Rodríguez, J., Bell, S.P. & Hochwagen, A. Mapping of meiotic single-stranded DNA reveals double-stranded-break hotspots near centromeres and telomeres. Curr. Biol. 17, 2003–2012 (2007).
17. Feng, W. et al. Genomic mapping of single-stranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. Nat. Cell Biol. 8, 148–155 (2006).
18. Feng, W., Bachant, J., Collingwood, D., Raghuraman, M.K. & Brewer, B.J. Centromere replication timing determines different forms of genomic instability in Saccharomyces cerevisiae checkpoint mutants during replication stress. Genetics 183, 1249–1260 (2009).
19. Leduc, F. et al. Genome-wide mapping of DNA strand breaks. PLoS ONE 6, e17353 (2011).
20. Dudley, D.D., Chaudhuri, J., Bassing, C.H. & Alt, F.W. Mechanism and control of V(DJ) recombination versus class switch recombination: similarities and differences. Adv. Immunol. 86, 43–112 (2005).
21. Shin, A.J., Chai, W., Shag, O.W. & Wright, W.E. Telomere-end processing the terminal nucleotides of human chromosomes. Mol. Cell 18, 131–138 (2005).
22. Palm, W. & Lange, T. How shelterin protects mammalian telomeres. Annu. Rev. Genet. 42, 301–334 (2008).
23. Casper, A.M., Nghiem, P., Arlt, M.F. & Glover, T.W. ATR regulates fragile site stability. Cell 111, 779–789 (2002).
24. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57, 289–300 (1995).
25. Smit, A. & Hubley, R. RepeatMasker Open v3.0 (http://www.repeatmasker.org/) (Institute for Systems Biology, Seattle, 1996–2004).
26. Dunkin, S.G. & Glover, T.W. Chromosome fragile sites. Annu. Rev. Genet. 41, 169–192 (2007).
27. Zhang, H. & Feudenberg, C.H. An AT-rich sequence in human common fragile site FRA1D causes fork stalling and chromosomal breakage in S. cerevisiae. Mol. Cell 27, 367–379 (2007).
28. Kim, N. & Jinks-Robertson, S. Transcription as a source of genome instability. Nat. Rev. Genet. 13, 204–214 (2012).
29. Helmich, A., Ballarino, M. & Tora, L. Collisions between replication and transcription complexes cause fragile site instability at the longest human genes. Mol. Cell 44, 966–977 (2011).
30. Halazonetis, T.D., Gorgoulis, V.G. & Bartek, J. An oncogenic-induced DNA damage model for cancer development. Science 319, 1352–1355 (2008).
31. Negrini, S., Gorgoulis, V.G. & Halazonetis, T.D. Genomic instability—an evolving hallmark of cancer. Nat. Rev. Mol. Cell Biol. 11, 220–228 (2010).
32. De, S. & Michor, F. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. Nat. Biotechnol. 29, 1103–1108 (2011).
33. Futreal, P.A. et al. A census of human cancer genes. Nat. Rev. Cancer 4, 177–183 (2004).
34. Santarius, T., Shipley, J., Brewer, D., Stratton, M.R. & Cooper, C.S. A census of amplified and overexpressed human cancer genes. Nat. Rev. Cancer 10, 59–64 (2010).
35. Ng, S.B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461, 272–276 (2009).
36. Altschuler, D. et al. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature 407, 513–516 (2000).
ONLINE METHODS

Cells, reagents and immunocytofluorescence. Mice experiments were approved by the Animal Care and Use Committee (ACUC) at Children’s Hospital, Boston (protocol number 13-01-2295). To obtain primary mouse single-cell suspensions, we squeezed testes and spleens from C57BL/6J mice between two microscope slides in a Petri dish filled with trypsin. We flushed bone marrow out of femurs and tibias from the same animals using a syringe filled with trypsin. We purified and activated B lymphocytes as previously described37. Prior to fixation for BLESS, we removed dead cells using a Ficoll gradient. We filtered cell suspensions through MACS Pre-Separation filters (Miltenyi Biotec) and then fixed them according to the BLESS protocol. We obtained IMR90 primary fibroblasts and HeLa cells from ATCC and U2OS_DRH-1 cells from Y. Shiloh (Tel Aviv University). A.J. Pierce (University of Kentucky) kindly communicated details on the construction of U2OS_DRH-1 cells. We transfected pcBAS-I-SceI and pCAGGs plasmids (kindly donated by Y. Shiloh) into U2OS_DRH-1 cells using Fugene (Roche) following the manufacturer’s instructions. After BLESS, we cloned gDNA fragments into pEFP-C1 (BD Biosciences). We applied aphidicolin (Sigma) onto cells at 0.4 µM for 18 h and nocodazostatin (Sigma) at 200 ng/ml for 45 min. We obtained oligonucleotide linkers from Sigma and annealed them in 1× T4 ligase buffer (NEB). Linkers and primers used are listed in Supplementary Table 5. We visualized γH2A.X foci by immunocytofluorescence (Millipore #05-636) and counted them as previously described38.

Breaks labeling, enrichment on streptavidin and sequencing (BLESS). A detailed step-by-step protocol to perform BLESS can be found on the supporting web page (http://www.breakome.eu/). Briefly, to prepare purified nuclei for in situ ligation, we fixed 5 million cells as single-cell suspensions in growth medium with 2% formaldehyde for 30 min at room temperature and then washed them once in ice-cold 1× PBS. To prepare single-nucleus suspensions, we first lysed fixed cells in a buffer containing 10 mM Tris-HCl, 10 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.2% NP-40, pH 8, for 90 min at 4 °C and then in a buffer containing 10 mM Tris-HCl, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 0.3% SDS, pH 8, for 45 min at 37 °C. We resuspended lysed cells in 1× NEBuffer 2 (NEB) supplemented with 0.1% Triton X-100 and proteinase K at 100 µg/ml final concentration. We mildly rotated cells for a short time at 37 °C (8 min for HeLa, 4 min for mouse B lymphocytes), after which we transferred them onto ice. We quenched proteinase K by adding an equal volume of buffer supplemented with PMSF. We washed purified nuclei twice in 1× NEBuffer 2 supplemented with 0.1% Triton X-100 and then once in blunting buffer (NEB) supplemented with 100 µg/ml BSA. We performed blunting using the Quick Blunting kit (NEB) according to the manufacturer’s instructions in a final volume of 100 µl for 45 min at room temperature. Afterwards, we washed nuclei twice in 1× NEBuffer 2 supplemented with 0.1% Triton X-100, once in 1× T4 ligase buffer supplemented with 0.1% Triton X-100 and once in 1× T4 ligase buffer. We performed in situ ligation for 18–20 h at 16 °C in 25 µl final volume using 1.5 µl of T4 ligase (NEB) and 5 µl of 10 µM proximal linker previously annealed in 1× ligase buffer. After ligation, we washed nuclei three times in a high-salt buffer (W&B) containing 5 mM Tris-HCl, 1 mM EDTA, 1 M NaCl, pH 7.5, supplemented with 0.1% Triton X-100. Afterwards, we extracted gDNA by incubating nuclei in 1× NEBuffer 2 with 0.5% Triton X-100 and proteinase K at 200 µg/ml final concentration for 1 h, shaking at 65 °C, followed by isopropanol-ethanol purification. We fragmented purified gDNA for 18–20 h at 37 °C using 6 U of HaeIII (NEB) per million cells fixed.

To capture labeled DSBs, we rotated 20 µg of purified gDNA with 5 µl of Dynabeads MyOne C1 (Invitrogen) in W&B buffer supplemented with 0.1% Triton X-100 for 30 min at 4 °C. Afterwards, we washed beads three times in W&B buffer supplemented with 0.1% Triton X-100 and then resuspended them in 37 µl of 1× T4 ligase buffer. We added 10 µl of distal linker, previously annealed at 10 µM in 1× ligase buffer, and 3 µl of T4 ligase to the beads, and the reaction was carried out for 16–18 h at 16 °C. After distal linker ligation, we washed beads twice in W&B buffer supplemented with 0.1% Triton X-100 at room temperature and then digested captured fragments with I-Sce (NEB) in 25 µl final volume for 4 h at 37 °C. Afterwards, we centrifuged beads and stored the supernatant at −20 °C until PCR was performed.

We used the entire volume of supernatant after I-Sce digestion to prepare multiple PCR amplification reactions (5 µl of supernatant per reaction) using Phusion polymerase (NEB) and the appropriate primer pair depending on the downstream sequencing platform (Supplementary Table 5). We performed 18 amplification cycles using conditions recommended by the manufacturer and T₀ = 55 °C. To remove unused primers, we purified PCR products in gel using the DNA Gel Extraction kit (Qiagen). Before Illumina library preparation, we digested purified PCR products with XhoI (NEB) to cleave terminal I-SceI sequences derived from linkers and again gel-purified them.

Next-generation sequencing. Sequencing was either outsourced (imaGenes GmbH, Berlin, Germany; and ServiceXS, Leiden, The Netherlands) or performed in-house as summarized in Supplementary Table 1. We prepared samples for Roche 454 sequencing using indexing barcode-containing primers during the PCR step in BLESS (Supplementary Table 5). We purified PCR products of sizes between 300 and 800 nt in gel and analyzed them on the 2100 Bioanalyzer (Agilent) before sequencing. For BLESS Illumina library preparation, we used the TruSeq DNA sample preparation kit v2 (Illumina) without DNA fragmentation and library size selection. For gDNA sequencing, we sheared gDNA with Covaris S220 AFA (Covaris) according to the manufacturer’s instructions before Illumina library preparation. We assessed library quality and quantity on the 2100 Bioanalyzer (Agilent) using the High Sensitivity DNA Kit (Agilent) and by qPCR using the Kapa Library Quantification Kit (Kapa Biosystems). We generated clusters on the Illumina flow cell using the automatic cBot station and the TruSeq PE Cluster Kit v3-cBot-HS. We carried out sequencing by synthesis on Illumina HiSeq 2000 using the TruSeq SBS Kit v3-HS chemistry.

ChIP and qPCR. We performed ChIP assays as previously described39, with minor modifications. We purified immunoprecipitated and input DNA with phenol-chloroform and analyzed it with real-time qPCR using primers listed in Supplementary Table 6. We compared the amount of DNA captured in untreated (C) vs. aphidicolin-treated (A) HeLa cells by qPCR followed by
data analysis according to the ΔCt method, using the Ct values obtained for each primer pair in sample C1 as a reference for the Ct values obtained for the same primer pair in sample A1. We used three technical replicates for each sample. Primers are listed in Supplementary Table 7.

**Computational analyses. Roche 454 data.** We analyzed Roche 454 data using previously described scripts developed in the Chiarle lab adapted for BLESS linkers. Briefly, we aligned sequences to the mouse reference genome (GRCm38/mm10) using BLAT and subsequently filtered them for the BLESS proximal linker using BLAST. We further processed filtered reads to remove PCR repeats (including repeats slightly divergent owing to sequencing errors), invalid alignments (including alignment scores <30, reads with multiple alignments having a score difference <4 and alignments having 10-nt gaps) and linker ligation artifacts (for example, random HaeIII restriction sites ligated to the proximal linker).

**Illumina data.** We analyzed Illumina data using the Instant-seq software suite developed in the Rowicka lab. We filtered reads from FASTQ files, requiring a Phred score of ≥20 for every base, and trimmed them at the point where the Phred score of an examined base fell below 20. We retained all reads with length ≥34 nt as high-quality filtered reads and scanned them for the presence of the exact proximal or distal barcode. After removal of barcodes, we aligned reads ≥23 nt to the GRCh37/hg19 assembly of the human genome. We retained only sequences mapping without mismatches to unique (U0) or multiple (R0) positions.

To identify aphidicolin-sensitive regions (ASRs), we compared the number of reads in A and C samples using windows with a constant number of mappable bases to account for the variation in mappability along the human genome and to produce more statistically robust comparisons of the number of reads between different windows. For comparison, we also used windows of constant length to demonstrate that our approach is not biased toward detecting ASRs in repetitive regions (Supplementary Fig. 7). Using a precomputed mappability map corresponding to 45-nt reads, we moved windows of chosen mappable length across each chromosom and calculated enrichment values according to the hypergeometric probability distribution. The parameters for calculating hypergeometric P values were the following: (i) total number of mapped reads in experiments A and C; (ii) total number of mapped reads in experiment A; (iii) total number of mapped reads in the sliding window in experiments A and C; and (iv) total number of mapped reads in the sliding window in experiment A only. We computed final P values using the Benjamini-Hochberg correction for multiple hypothesis testing. We first analyzed individual sample replicates (A1 vs. C1; A2 vs. C2; A3 vs. C3; A4 vs. C4) and found that ASRs were highly correlated (Supplementary Table 2). Therefore, for all subsequent analyses, we pooled samples so that C = (C1 + C2 + C3 + C4) and A = (A1 + A2 + A3 + A4). To account for copy-number variation effects due to karyotype or aphidicolin treatment, we sequenced gDNA derived from normal human fibroblasts (g-F), HeLa (g-C) and HeLa treated with aphidicolin (g-A) and mapped it to the human reference genome (GRCh37/hg19) (Supplementary Table 1). To correct for CNV and aphidicolin effect, using windows with a constant number of mappable bases, we first calculated the expected number of reads in each window on the basis of the g-F sample and then compared it to normalized reads obtained from g-C and g-A samples (Supplementary Fig. 4b).

We compared the number of reads in A versus C samples. If the ratio of reads was higher or lower than the expected number, we computed the P value of the corresponding enrichment or depletion. P values were then corrected for multiple hypothesis testing using Benjamini-Hochberg correction, and significantly enriched windows (P ≤ 0.05) were finally annotated as high-confidence, karyotype- and aphidicolin-corrected ASRs.

To compute gene sensitivity to aphidicolin, we obtained the start and end coordinates of each gene annotated in GencodeV12 by combining all annotated transcripts. We analyzed each gene by using the same approach described above for ASRs, using the start-end coordinates of the gene as the start-end coordinates of genomic windows. Genes that did not have any mappable bases within the boundaries of the window were not analyzed.

We inferred the relationship between aphidicolin sensitivity and gene length by computing rank correlation. We also computed the overlap between genes and ASRs, by agglomerating results from 50 different ASR maps with resolutions in the range of 10–60 mappable kilobases. For each window length, we first calculated the percentage overlap of a gene with ASRs and then averaged percentages over all lengths. To assess the significance of overlap enrichment, we computed the distribution of averaged overlap for each gene. We calculated the exact distribution under the assumption that the aphidicolin sensitivity of each window is independently assigned, with probability depending on its length (a different percentage of windows was aphidicolin-sensitive for different window sizes as shown in Supplementary Fig. 4a).

We computed the exact distribution step by step for consecutive window lengths using dynamic programming. Based on this distribution, we calculated for each gene the P value of its overlap enrichment and then applied Benjamini-Hochberg multiple hypotheses testing correction to the whole gene lists. We compared ASRs enrichment in cancer-associated genes vs. all genes using both the Kolmogorov-Smirnov and hypergeometric tests, obtaining the same result (P = 0.04). Finally, we identified genes that have transcripts with S′ end within a 2-Mb vicinity of an ASR center using the Gencode annotation of genes. We extracted locations of all transcripts, and for each ASR we identified transcripts in its proximity and calculated the distance from the center of the ASR to the transcription start site. We analyzed the list of transcripts within 2 Mb of an ASR and reported the corresponding gene along with the exact distance and ASR center location (within or outside). This list was binned at various intervals of distances. Finally, we analyzed the list of transcripts in the vicinity of ASRs and output a list of gene groups binned according to the distance from the center of the ASR to the closest transcription start site (Fig. 4c). We computed statistics for differences between groups using the hypergeometric probability distribution.

For biological characterization of ASRs, we created feature data sets from RepeatMasker and from the CpG Islands tracks of the UCSC Genome Browser and the genome-wide summary data of the Tumorscape portal. We obtained cancer-associated genes from the Sanger Institute’s Cancer Gene Census and when required, we mapped genome coordinates to the GRCh37/hg19 assembly of the human genome using liftOver. As repetitive regions tend to be underrepresented among uniquely mapped reads, we corrected for differences in mappability. To determine
whether a given genomic feature is enriched in ASRs, we computed the proportion of mappable nucleotides belonging to both the ASRs and the feature, as well as the proportion of ASRs among all the intervals considered. Next we performed 100,000 permutations of ASR assignments among the windows considered. Using these permutations, we calculated the empirical distribution of the ratio under the null hypothesis that the given feature and ASRs are independently distributed in the human genome. We used this distribution to estimate the *P* value for the feature enrichment inside ASRs. This method yielded a *P* value resolution of $10^{-5}$, which was too low for features particularly well correlated with fragility (Supplementary Table 3). In such cases, we analyzed empirical ratio distributions. We observed that the normal distribution fit well for long windows (48 mappable kilobases). For short windows (250 and 2,000 mappable nucleotides), the number of ASR mappable nucleotides within featured regions was almost always the integer multiple of the window length, and multiplication factors followed the geometric distribution. Thus, we analytically determined *P* values according to these distributions.

37. Chiarle, R. *et al.* Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* **147**, 107–119 (2011).
38. Crosetto, N. *et al.* Human Wrnlp1 is localized in replication factories in a ubiquitin-binding zinc finger-dependent manner. *J. Biol. Chem.* **283**, 35173–35185 (2008).
39. Tyteca, S., Vandromme, M., Legube, G., Chevillard-Briet, M. & Trouche, D. Tip60 and p400 are both required for UV-induced apoptosis but play antagonistic roles in cell cycle progression. *EMBO J.* **25**, 1680–1689 (2006).
40. Schmittgen, T.D. & Livak, K.J. Analyzing real-time PCR data by the comparative $C_T$ method. *Nat. Protoc.* **3**, 1101–1108 (2008).
41. Fujita, P.A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.* **39**, D876–D882 (2011).
42. Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).